Analysis of Global and Local Optima of Regularized Quantile Regression in High Dimension: A Subgradient Approach

LAN WANG [a] AND XUMING HE[b]

University of Miami [a] and University of Michigan [b]

Abstract

Regularized quantile regression is a useful technique for analyzing heterogeneous data under potential heavy-tailed error contamination in high dimension. This paper provides a new analysis of the estimation/prediction error bounds of the global solution of L_1-regularized quantile regression and the local solutions of nonconvex regularized quantile regression when the number of covariates is greater than the sample size. Our results build upon and significantly generalize the earlier work in the literature. For heavy-tailed error distributions and a general class of design matrices, the least squares-based LASSO cannot achieve the near-oracle rate derived under the normality assumption no matter the choice of the tuning parameter. In contrast, we establish for quantile regression with the LASSO-type penalty the near-oracle estimation error rate for a broad class of models under conditions weaker than those in the literature. For nonconvex regularized quantile regression, we establish the novel results that all local optima within a feasible region share the desirable minimax error bounds for parameter estimation, as well for the quantile prediction errors. Our analysis applies to not just the hard sparsity setting commonly used in the literature but also to models under soft sparsity assumption which permits many covariates to have small effects. Our approach relies on a unified characterization of the global/local solutions of regularized quantile regression via subgradients using a generalized KKT condition. The theory of the

[1]Lan Wang is Professor, Department of Management Science, University of Miami. Emails: lanwang@mbs.miami.edu. Xuming He is H. C. Carver Collegiate Professor, Department of Statistics, University of Michigan. Email: xmhe@umich.edu. Wang and He's research is supported by NSF FRGMS-1952373.

paper establishes a key property of the subdifferential of the quantile loss function in high dimensions, which is of independent interest for analyzing other high-dimensional nonsmooth problems.

Keywords: Estimation error, Quantile regression, High-dimensional data, L1 penalty, Nonconvex penalty

1 Introduction

The semiparametric technique of quantile regression (QR) provides a useful alternative to least-squares regression and has been widely applied to analyze data arising in economics and finance, since its introduction in the seminal paper of Koenker and Bassett (1978). For example, a low quantile of the return distribution of an investment portfolio provides an assessment of risk commonly known as Value-at-Risk. Buchinsky et al. (1994), Chamberlain (1994), Buchinsky (1998), Abadie et al. (2002), Horowitz and Spokoiny (2002), Angrist et al. (2006), Firpo et al. (2009), Galvao et al. (2013), Arellano and Bonhomme (2017), Graham et al. (2018), among others, employed quantile regression to study the wage distribution. See also Horowitz and Lee (2005), Koenker and Xiao (2006), Chernozhukov and Fernández-Val (2011), Chernozhukov et al. (2013), Fitzenberger et al. (2013), Linton and Whang (2004), Su and Hoshino (2016), Koenker (2017) and Koenker et al. (2017) for other interesting applications of quantile regression in economics. Quantile regression helps characterize the entire conditional distribution and often leads to discoveries of interesting features of the data that would otherwise be imperceptible. It also has the appealing property of being robust to heavy-tailed error distributions. By contrast, the L_1-regularized least squares regression, to be called LS-LASSO throughout the paper (Tibshirani (1996)), is known to be vulnerable to heavy-tailed errors.

Let Y be a random variable and $X = (x_1, \ldots, x_p)^T$ be a p-dimensional vector of covariates. A linear quantile regression model takes the form

$$Y = X^T\beta^* + \epsilon, \qquad P(\epsilon \leq 0|X) = \tau, \text{ for some } 0 < \tau < 1, \tag{1}$$

where the error distribution of ϵ is generally heteroscedastic, and $\beta^* = (\beta_1^*, \beta_2^*, ..., \beta_p^*)^T$ is the unknown parameter vector. In this formulation, both ϵ and β depend on the quantile level τ of interest but we ignore such dependence in notation for simplicity. Model (1) implies that $Q_{Y|X}(\tau) = X^T\beta^*$, where $Q_{Y|X}(\tau) = \inf\{t : F_{Y|X}(t) \geq \tau\}$ is the τth conditional quantile of Y given X. We are interested in estimating β^* in the setting where the number of covariates is much larger than the sample size.

This paper develops a useful technique to study high-dimensional quantile regression in a general framework under a set of lean assumptions. Our theory relies on establishing a key property of the subdifferential of the quantile loss function in high dimension. Let $S_n(\beta)$ be any subgradient of the sample quantile loss function (with more details in Section 2.3), and $\Delta = \beta - \beta^*$. We shall show that there exist positive constants k_0^* and c^* which do not depend on n or p such that

$$\langle S_n(\beta) - S_n(\beta^*), \Delta \rangle \geq k_0^* ||\Delta||_2^2 - c^* \sqrt{\frac{\log p}{n}} ||\Delta||_1,$$

uniformly on $\{||\Delta||_2 \leq 1\} \cap \mathbb{C}$, with high probability, where the set $\mathbb{C}$, to be made clearer later in the paper, depends on the specific estimation method under study and the sparsity of the true parameter β^*.

The subgradient approach leads to a unified analysis of both the global solution of the L_1-regularized quantile regression (QR-LASSO) and the local solutions of nonconvex regularized quantile regression (QR-NCP) under a set of mild assumptions, and allowing for a more general sparsity pattern of β^*. We include both QR-LASSO and QR-NCP in the paper because both types of penalty functions are of substantial interest in the literature and in practice. QR-LASSO is computationally convenient due to the convexity in the objective function, whereas QR-NCP can help alleviate the bias due to the over-penalization of the L_1-penalty. For regularized least squares regression and generalized linear models in high dimension, an equivalent property of restricted strong convexity has been shown to play a fundamental role. However, such results are available only for differential loss functions. The gradient function of the quantile loss is not even Lipschitz continuous, which leads to substantial technical challenges. Our proof involves a novel construction of a Lipschitz continuous lower bound and the use of modern empirical processes techniques.

In the classical setting where the number of covariates is not large, Wang et al. (2007), Li and Zhu (2008), Zou and Yuan (2008), Wu and Liu (2009), Shows et al. (2010), Kai et al. (2011), Wagener et al. (2012), Wang et al. (2013a), Chen et al. (2019a), among others, investigated regularized quantile regression for variable selection. Several authors have recently investigated quantile regression in high dimension. Belloni and Chernozhukov (2011) was among the first to rigorously establish estimation error bounds for QR-LASSO, see also Wang (2013). More recently, Park et al. (2017) investigated multiple quantile regression with high dimensional covariates, Lee et al. (2018) studied high-dimensional QR-LASSO with a change point, Harding and Lamarche (2018) investigated QR-LASSO for panel data, Chen et al. (2019b) explored quantile regression for big data under memory constraint. Moreover, adaptively weighted L_1-regularized or nonconvex regularized quantile regression has been considered for better variable selection in various settings; see Bradic et al. (2011), Wang et al. (2012), Fan et al. (2014), Zheng et al. (2015), among others. High-dimensional semi-parametric quantile regression have been investigated by Tang et al. (2013), Sherwood et al. (2016), Zhong et al. (2016), Fan and Lian (2018), Lv et al. (2018), Honda et al. (2019), among others.

Inspired by the recent work, this paper makes several contributions to the fundamental theory of quantile regression in the asymptotic regime where $p = \exp(n^{d_0})$ for some constant $d_0 > 0$, i.e., the ultrahigh dimensional regime.

- We show that QR-LASSO enjoys the near-optimal estimation error rate under a set of lean assumptions. Our theory permits a rich class of error distributions as well as a general class of random design matrices without requiring the *nonlinear eigenvalue condition*. The estimation error rates are established under not only the popular hard sparsity setting, but also a more relaxed soft sparsity assumption which permits many covariates to have small effects.

- For QR-NCP, we show that all local minima within a feasible region have the desirable error bounds and achieve the minimax error rate of estimation. These new results fill an important theoretical gap in the literature, because the global minimizers for nonconvex objective functions are not always numerically obtained or verifiable in practice.

- We derive the quantile prediction error rate by a general characterization of the prediction error based on subgradients.

Our results demonstrate that QR-LASSO enjoys near-oracle estimation error rates for a much richer class of error distributions than LS-LASSO does. The theory relies on conditions generally weaker than those in the current literature for LS-LASSO and QR-LASSO. Our analysis of the local minima for QR-NCP is new. Earlier result in the literature on QR-NCP are mostly focused on the existence of a local solution with good statistical property. In contrast, this paper shows that any local solution satisfying the first-order condition within a given radius of the true value has desirable statistical accuracy. This result has important implications for the use of QR-NCP as an estimator for the quantile regression coefficient or as an initial value for inference (see Section 6 for more discussions). This substantially generalizes the results in Loh and Wainwright (2015), Mei et al. (2018) and Elsener and van de Geer (2018) on the properties of local minima for differentiable loss functions. It is worth noting that the statistical properties of local solutions are of broader interest. Even in low dimension, algorithms for many nonlinear problems (e.g., nonlinear GMM) only guarantee first-order solutions (e.g., stationary point). Moreover, our subgradient approach is different from the techniques commonly used in the literature of high-dimensional quantile regression and is of significant independent interest. The proposed technique can potentially be applied to a large class of high-dimensional nonsmooth problems, for instance, classification based on the hinge loss.

The rest of the paper is organized as follows. Section 2 introduces the background, provides an example where LS-LASSO has suboptimal performance, and discusses a general characterization of the solution of regularized quantile regression based on subgradients. Section 3 presents the main theory on the estimation error bounds for QR-LASSO and QR-SCAD in ultra-high dimension. Section 4 studies the quantile prediction error bound. Section 5 reports results from a Monte Carlo study. Section 6 concludes the paper with additional remarks. The appendices contain detailed technical arguments.

Notation. In this paper, for any vector $v = (v_1, \ldots, v_p)^T \in \mathbb{R}^p$, $||v||_2 = \sqrt{\sum_{i=1}^p v_i^2}$ denotes its L_2 norm, $||v||_1 = \sum_{i=1}^p |v_i|$ denotes its L_1-norm, and $||v||_\infty = \max_{1\le i\le p} |v_i|$ denotes its L_∞ norm. Given an arbitrary index set $S \in \{1, \ldots, p\}$, let v_S denote the subvector of v containing the elements whose indice are in S. For a real symmetric matrix M, $\lambda_{max}(M)$ denotes its largest eigenvalue. For sequences of real positive numbers a_n an b_n, $a_n \sim b_n$ means $c_1 \le a_n/b_n \le c_2$ for some positive constants c_1 and c_2. In this paper, a number is referred to as "a constant" if it does not depend on (n, p) (but it is allowed to depend on some features (e.g. variance proxy) of the probability distributions of X_i and ϵ_i). The indicator function of an event A, denoted by $\mathrm{I}(A)$, takes the value one if A happens and zero otherwise.

2 Preliminaries

In this section, we briefly introduce regularized quantile regression with convex and nonconvex penalties, under the hard sparisty or soft sparsity assumption. We then elaborate on how to use subgradiate to characterize the regularized quantile regression solution, global or local, with a general penalty function.

2.1 Background

Consider a random sample $\{Y_i, X_i\}_{i=1}^n$ satisfying model (1), where $X_i = (x_{i1}, \ldots, x_{ip})'$. To explicitly incorporate the intercept term, we assume $x_{i1} = 1$ for all i, and correspondingly $\beta^* = (\beta_1^*, \beta_-^{*T})^T$ where $\beta_-^* = (\beta_2^*, \ldots, \beta_p^*)^T$.

To avoid overfitting in the setting $p \gg n$, we consider estimating β^* by regularized quantile regression:

$$\widehat{\beta} \equiv \widehat{\beta}(\lambda) \equiv \underset{\beta\in\mathbb{R}^p}{\arg\min} \Big\{ n^{-1} \sum_{i=1}^n \rho_\tau(Y_i - X_i^T \beta) + \sum_{j=2}^p p_{\lambda_n}(|\beta_j|) \Big\}, \tag{2}$$

where $\lambda = \lambda_n$ is positive tuning parameter, $\rho_\tau(u) = u\{\tau - \mathrm{I}(u < 0)\}$ is the quantile loss function, $\beta = (\beta_1, \beta_-^T)^T \in \mathbb{R}^p$, and $p_{\lambda_n}(\cdot)$ denotes a penalty function with tuning parameter

λ_n which controls the complexity of the solution. A larger value of λ_n encourages sparser solutions.

QR-LASSO adopts the popular L_1-penalty for which $p_{\lambda_n}(|\beta_j|) = \lambda_n|\beta_j|$. It is computationally convenient due to its convex structure. Alternatively, one may use a nonconvex penalty function, which can alleviate the bias associated with the L_1-penalty. Two popular choices of nonconvex penalty functions are the SCAD penalty function (Fan and Li (2001)) and the MCP penalty function (Zhang (2010)). The SCAD penalty function is defined by

$$\begin{aligned} p_\lambda(|\beta|) &= \lambda|\beta| I(0 \le |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} I(\lambda \le |\beta| \le a\lambda) \\ &\quad + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda), \text{ for some } \ a > 2. \end{aligned}$$

The MCP function has the form

$$p_\lambda(|\beta|) = \lambda\Big(|\beta| - \frac{\beta^2}{2a\lambda}\Big) I(0 \le |\beta| < a\lambda) + \frac{a\lambda^2}{2} I(|\beta| \ge a\lambda), \quad \text{for some } \ a > 1.$$

Most of the existing literature on high-dimensional quantile regression assumes that β^* satisfies a hard sparsity constraint, specifically,

$$\beta^* \in \mathbb{B}_0(s) = \Big\{\beta \in \mathbb{R}^p : \sum_{j=2}^p \mathrm{I}(\beta_j \neq 0) \le s - 1\Big\}, \tag{3}$$

for some positive constant $2 \le s \ll n$, where $\mathrm{I}(\cdot)$ denotes the indicator function. Hence, $||\beta^*||_0 \le s$ and most of the components in β^* are exactly zero. As is customary for a regression model in the classical setting, the intercept term is always included in the model. The sparsity constraints are thus imposed on the slope components of β^*. Note that this is a subtle difference from high-dimensional least squares regression where the intercept is generally taken as zero, which can be done for mean regression without loss of generality by centering both the response variable and the regressors, but we cannot do the same for quantile regression.

The hard sparsity constraint may be overly restrictive for some applications where many weak signals, rather than just a few strong signals, are likely to be relevant. This paper also considers a more relaxed sparsity constraint, which allows β^* to have many smallish nonzero coefficients. More specifically, the soft sparsity constraint assumes

$$\beta^* \in \mathbb{B}_1(R) = \left\{\beta \in \mathbb{R}^p : \sum_{j=2}^{p} |\beta_j| \leq R\right\} \tag{4}$$

for some positive number R, which may depend on the sample size. In (4), instead of using the L_1-norm, we may also use the L_q norm for some $0 < q < 1$. The results of the paper would still hold under minor modifications. It is worth noting that both $\mathbb{B}_0(s)$ and $\mathbb{B}_1(R)$ depend on the quantile level τ of interest. Further discussions on the identification of the population parameter β_0 in the high-dimensional setting are given in Appendix A.

2.2 Characterizing the solution for regularized quantile regression

We now present a unified characterization of the regularized quantile regression estimator, including the global solution of QR-LASSO and the local solutions of QR-NCP, based on a generalized KKT condition for the convex difference problem, characterized by subgradients.

A subgradient of a convex function $g(\beta)$ at β_1 is any vector $\xi \in \mathbb{R}^p$ such that $g(\beta_2) \geq g(\beta_1) + \xi^T(\beta_2 - \beta_1)$ for all β_2. The subdifferential of $g(\beta)$ at β_1, denoted by $\partial g(\beta_1)$, consists of all the subgradients of $g(\beta)$ at β_1.

Let

$$Q_n(\beta) = n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^T \beta) \tag{5}$$

denote the empirical quantile loss function. The sub-differential $\partial Q_n(\beta)$ is the collection of vectors $S_n(\beta) = (S_{n1}(\beta), ..., S_{np}(\beta))^T$, where for $j = 1, ..., p$,

$$\begin{aligned} S_{nj}(\beta) \quad = \quad & -\tau n^{-1} \sum_{i=1}^{n} x_{ij} \mathrm{I}(Y_i - X_i^T \beta > 0) \\ & +(1-\tau) n^{-1} \sum_{i=1}^{n} x_{ij} \mathrm{I}(Y_i - X_i^T \beta < 0) - n^{-1} \sum_{i=1}^{n} x_{ij} v_i \end{aligned} \tag{6}$$

and

$$v_i = \begin{cases} 0, & \text{if } Y_i - X_i^T\beta = 0 \\ \in [\tau - 1, \tau], & \text{otherwise} \end{cases}$$

Let $L_n(\beta) = Q_n(\beta) + \sum_{j=2}^p p_{\lambda_n}(|\beta_j|)$ be the regularized quantile objective function $L_n(\beta)$ in (2). We observe that for a general class of penalty functions, $L_n(\beta)$ can be written as the difference of two convex functions in β:

$$L_n(\beta) = \tilde{L}_n(\beta) - H(\beta),$$

where $\tilde{L}_n(\beta) = Q_n(\beta) + \lambda \sum_{j=2}^p |\beta_j|$ and $H(\beta) = \sum_{j=2}^p h_\lambda(\beta_j)$, where $h_\lambda(.)$ is differentiable. In the case of LASSO-penalty, $h_\lambda(\beta_j) = 0$, for $j = 2, \ldots, p$, thus $\tilde{L}_n(\beta)$ coincides with $L_n(\beta)$. For the nonconvex SCAD penalty,

$$\begin{aligned} h_\lambda(\beta_j) &= \Big[(\beta_j^2 - 2\lambda|\beta_j| + \lambda^2)/(2(a-1))\Big] I\big(\lambda \le |\beta_j| \le a\lambda\big) \\ &\quad + \Big[\lambda|\beta_j| - (a+1)\lambda^2/2\Big] I\big(|\beta_j| > a\lambda\big); \end{aligned}$$

while for the nonconvex MCP function,

$$h_\lambda(\beta_j) = \Big[\beta_j^2/(2a)\Big] I\big(0 \le |\beta_j| < a\lambda\big) + \Big[\lambda|\beta_j| - a\lambda^2/2\Big] I\big(|\beta_j| \ge a\lambda\big).$$

For the above convex difference optimization problem, an extension of the KKT condition was given in Tao and An (1997), which implies that the solution $\widehat{\beta}$ of (2), global or local, satisfies the following necessary condition:

$$\nabla \tilde{L}_n(\widehat{\beta}) - H'(\widehat{\beta}) = 0, \tag{7}$$

where $\nabla \tilde{L}_n(\widehat{\beta})$ denotes some (not necessarily any) subgradient in the subdifferential of $\nabla \tilde{L}_n$ being evaluated at $\widehat{\beta}$, and $H'(\widehat{\beta}) = (0, h'_\lambda(\widehat{\beta}_2), \ldots, h'_\lambda(\widehat{\beta}_p))^T$. Note that $\tilde{L}_n(\beta)$ is the sum of two convex functions. In this paper, we consider stationary points satisfying (7), which by the lemma below, include the solutions of interest.

Lemma 1 *Let $\widehat{\beta}$ be the global minimum of QR-LASSO or the local minimum of QR-NCP. Then there exists a subgradient $S_n(\beta) \in \partial Q_n(\beta)$ such that*

$$S_n(\widehat{\beta}) + \lambda sgn(\widehat{\beta}) - H'(\widehat{\beta}) = 0, \tag{8}$$

where $sgn(\widehat{\beta}) = (0, sgn(\widehat{\beta}_2), \ldots, sgn(\widehat{\beta}_p))^T$, and the sign function $sgn(t) = 1$ if $t > 0$; $= -1$ if $t < 0$; and takes its value in $[-1, 1]$ if $t = 0$, $j = 1, \ldots, p$.

Consider next a particular subgradient $\widetilde{S}_n$ in $\partial Q_n(\beta)$, given by

$$\widetilde{S}_n = n^{-1} \sum_{i=1}^{n} X_i \xi_i, \tag{9}$$

where $\xi_i = \mathrm{I}(\epsilon_i < 0) - \tau$. regularized quantile regression usually selects λ such that the event

$$\Lambda_n = \left\{ \lambda \geq c_0 ||\widetilde{S}_n||_\infty \right\} \tag{10}$$

happens with a large probability for some positive constant $c_0 > 1$. Such a penalty was also considered in Belloni and Chernozhukov (2011) for L_1-regularized quantile regression, see also Kato (2011) for an extension to the group Lasso setting. The above choice of tuning parameter is motivated by the general principal of tuning parameter selection in regularized least-squares regression (Bickel et al. (2009)) and the KKT condition for a general convex difference problem. Following the choice for LS-LASSO (Bickel et al. (2009)), we take $c_0 = 2$ in the subsequent analysis.

3 Main theory

In this section, we provide details on the theoretical properties of regularized quantile regression estimator $\widehat{\beta}$. For QR-LASSO, $\widehat{\beta}$ denotes the global solution defined in (2), which also satisfies (7). For QR-NCP, $\widehat{\beta}$ denotes any local solution satisfying (7).

3.1 Geometric structure of regularized quantile regression estimator

Under the hard sparsity condition, the QR-LASSO estimator is known to lie in a cone-shaped set with high probability. Let $\widehat{v} = \widehat{\beta} - \beta^*$. Here, we go beyond the setting of the L_1-penalty and hard sparsity to characterize the geometric structure of $\widehat{v}$.

Let $S_- = \{j : \beta_j^* \neq 0,\ 2 \leq j \leq p\}$ and $S = S_- \cup \{1\}$. Let $S_{-a} = \{j : |\beta_j^*| > a, 2 \leq j \leq p\}$, where a is a small positive thresholding parameter to be discussed later. Let $S_a = S_{-a} \cup \{1\}$. The cardinality $||S||_0 = s$ denotes the sparsity size under the hard sparsity condition. Under the soft sparsity assumption, s can be much larger than n.

Let

$$
\begin{aligned}
\Gamma_H &= \big\{v \in \mathbb{R}^p : ||v_{S^c}||_1 \leq 3||v_S||_1\big\}, & (11)\\
\Gamma_W &= \big\{v \in \mathbb{R}^p : ||v_{S_a^c}||_1 \leq 3||v_{S_a}||_1 + 4||\beta^*_{S_a^c}||_1\big\}, & (12)\\
\widetilde{\Gamma}_H &= \big\{v \in \mathbb{R}^p : ||v_{A^c}||_1 \leq 3||v_A||_1\big\}, & (13)\\
\widetilde{\Gamma}_W &= \big\{v \in \mathbb{R}^p : ||v_{S_a^c}||_1 \leq 3||v_{S_a}||_1 + 2||\beta^*_{S_a^c}||_1\big\}, & (14)
\end{aligned}
$$

where v_A denotes the subvector of v that contains the s-largest elements of v. It is clear that Γ_H and $\widetilde{\Gamma}_H$ are cone shaped, but Γ_W and $\widetilde{\Gamma}_W$ are star-shaped. To see this, we simply note that if $v \in \Gamma_W$, then the whole line segment $\{t|t \in (0,1)\}$ is contained in Γ_W. Furthermore, in the definition of Γ_W or $\widetilde{\Gamma}_W$, a can be taken as any positive constant.

We use $f_i(t)$ to denote the conditional probability density function of ϵ_i given X_i, $i = 1, \ldots, n$. We also assume, without loss of generality, that the covariate $X_{-i} = (x_{i2}, \ldots, x_{ip})^T$ is a $(p-1)$-dimensional mean zero random vector, and $\Sigma = \mathrm{E}(X_i X_i^T)$ exists. Conditions (C1)-(C3) below constitute a set of basic assumptions for establishing the statistical properties of the regularized quantile regression estimator in high dimension.

Condition (C1). The conditional distribution of ϵ_i satisfies $P(\epsilon_i \leq 0|X_i) = \tau$, $i = 1, \ldots, n$. There exist positive constants m_0 and b_0 such that $\inf_{1\leq i\leq n} f_i(t) \geq m_0 > 0$, for all $|t| \leq b_0$.

Condition(C2). The matrix Σ satisfies $\lambda_{\max}(\Sigma) \leq k_u < \infty$, and

$$v^T \Sigma v \geq m_1 ||v||_2^2, \quad \text{for any} \quad v \in \mathbb{C}, \tag{15}$$

for some constant m_1, where for QR-LASSO, $\mathbb{C} = \Gamma_H$ under the hard sparsity assumption, and $\mathbb{C} = \Gamma_W$ under the soft sparsity assumption; while for QR-NCP, $\mathbb{C} = \mathbb{R}^p$.

Condition (C3). Let $\widehat{\sigma}_j^2 = n^{-1} \sum_{i=1}^n x_{ij}^2$. There exist a constant $m_x > 0$ and a positive sequence of numbers δ_n such that $P\big(\max_{1 \leq j \leq p} \widehat{\sigma}_j^2 \leq m_x\big) \geq 1 - \delta_n$, where $\delta \to 0$ as $n \to \infty$.

Remark 1. Condition (C1) imposes regularity conditions on the random error distributions, which allows for heteroscedastic error distributions and requires no moments. The constant b_0 in (C1) may depend on the probability distribution of X_i, as described in Lemma C.2 of Appendix C. A large class of heavy-tailed error distributions such as the Cauchy distribution satisfy condition (C1) . The restricted eigenvalue condition in (C2) is similar to chose imposed for regularized least-squares regression. For QR-LASSO, (15) is exactly the same as the restricted eigenvalue condition for LS-LASSO as the restriction sets Γ_H and Γ_W have the same forms as those for LS-LASSO. For QR-NCP, the requirement of $\mathbb{C} = \mathbb{R}^p$ amounts to assuming $\lambda_{\min}(\Sigma) \geq m_1 > 0$. However, if we restrict our attention to a sparse local solution, then this can be replaced by weaker sparse eigenvalue condition in Zhang (2010). Finally, condition (C3) is satisfied if the covariates have sub-Gaussian distributions. It can also be satisfied when some of the covariates do not have sub-Gaussian distributions. For example, if a small subset (say fixed size) of covariates only have finite second moments while others follow the sub-Gaussian distributions with bounded variance proxy, then (C3) still holds. Overall, the above set of conditions are similar to or weaker than those in the literature for high-dimensional quantile regression. Some detailed comparisons are given in Remark 3 of Section 3.3.

Lemma 2 *(QR-LASSO) Assume $\lambda = k_0\sqrt{\log p/n}$, where $k_0 \geq 4\sqrt{m_x}$ is a constant. Suppose conditions (C1) and (C3) are satisfied. Then with probability at least $1 - \delta_n - 2\exp(-\log p)$,* *(i) $\widehat{v} \in \Gamma_H$ under the hard sparsity assumption; and (ii) $\widehat{v} \in \Gamma_W$ under the soft sparsity assumption.*

For QR-LASSO, the geometric structure is a result of the convexity of the regularized quantile loss function. The first part of the Lemma 2 under the hard sparsity was observed in Belloni and Chernozhukov (2011); while the result under soft sparsity is new and is a generalization of Negahban et al. (2012). For QR-NCP, the geometric structure is less transparent. Instead, the structure is implicit in the derivation of the estimation error bound.

For QR-NCP, we have $\widehat{v} \in \widetilde{\Gamma}_H$ under the hard sparsity assumption and $\widehat{v} \in \widetilde{\Gamma}_W$ under the soft sparsity assumption with high probability. Due to the reliance on the conditions of later theorems, we refer to Corollary 1 in Section 3.3 for a full description of the results.

3.2 Properties of subgradients in high dimension

We first state a useful property for the subgradient $\widetilde{S}_n$ defined in (9). The following lemma gives a high probability bound for its supremum norm.

Lemma 3 *Suppose conditions (C1) and (C3) are satisfied. We have*

$$P\big(||\widetilde{S}_n||_\infty \leq 2\sqrt{m_x \log p/n}\big) \geq 1 - \delta_n - 2\exp(-\log p).$$

The lemma suggests that the event Λ_n defined in (10) occurs with a high probability for an appropriate choice of λ at the rate $\sqrt{\log p/n}$.

Theorem 1 below provides a core result for establishing error bounds for QR-LASSO and QR-NCP. It shows that a type of restricted convexity condition holds with high probability for any subgradient in the sudifferential of quantile loss.

Theorem 1 *Suppose conditions (C1)-(C3) are satisfied. There exist some positive constants a^*, c^*, a_1 and a_2, such that for any subgradient $S_n \in \partial Q_n(\beta)$,*

$$\langle S_n(\beta^* + \Delta) - S_n(\beta^*), \Delta \rangle \geq a^* ||\Delta||_2^2 - c^* \sqrt{\frac{\log p}{n}} ||\Delta||_1, \tag{16}$$

uniformly on $\{||\Delta||_2 \le 1\} \cap \mathbb{C}$*, holds with probability at least* $1 - \delta_n - a_1 \exp(-a_2 \log p)$*, where* $\delta_n \to 0$ *as* $n \to \infty$*.*

Remark 2. This theorem guarantees a restricted strong convexity (RSC) condition. Lemma C.4 in Appendix A shows that the uniform lower bound in the above theorem on $\{||\Delta||_2 \le 1\} \cap \mathbb{C}$ can be extended to $\{||\Delta||_2 > 1\} \cap \mathbb{C}$. Specifically, with probability at least $1 - \delta_n - a_1 \exp(-a_2 \log p)$, $\langle S_n(\beta) - S_n(\beta^*), \Delta \rangle \ge a^* ||\Delta||_2 - c^* \sqrt{\frac{\log p}{n}} ||\Delta||_1$, uniformly on $\{||\Delta||_2 > 1\} \cap \mathbb{C}$. The RSC conditions play a critical role in the recent literature on high-dimensional M-regression (Negahban et al. (2012), Loh and Wainwright (2015), among others). However, the existing literature has considered only smooth (second-order differential) loss functions. To our best knowledge, this is the first time, the RSC condition is established for a nonsmooth loss function, which is more technically challenging due to the fact the gradient function is not even Lipschitz continuous. Our proof is based on a novel construction of a Lipschitz continuous lower bound and the application of advanced empirical process theory techniques (e.g., peeling). Our approach can be applied to other nonsmooth high-dimensional problems and is of interest beyond quantile regression.

3.3 Near-oracle estimation error bound

This subsection derives the L_2-error and L_1-error bounds for the estimator $\widehat{\beta}$ under both the hard sparsity assumption and the soft sparsity assumption. It is worth emphasizing that the results here are nonasymptotic in the sense that the error bounds hold for any (n, p) satisfying the stated conditions. The theory allows the number of covariates p to grow at an exponential rate of the sample size n, often called the ultra-high dimensional setting.

Theorem 2 *(QR-LASSO) Suppose conditions (C1)-(C3) are satisfied. Let* $\lambda = k_0 \sqrt{\log p / n}$*, where* $k_0 \ge 4\sqrt{m_x}$ *is a constant. Let* a_1 *and* a_2 *be the positive constants given in Theorem 1. (i) (Hard-sparsity case) Assume* β^* *satisfies the hard sparsity assumption (3), with* $n > c' s \log p$*, where* $c' = 16(2k_0 + c^*)^2/(a^*)^2$*. Let* a_1 *and* a_2 *be the positive constants in Theorem 1. Then with probability at least* $1 - 4\delta_n - 4\exp(-\log p) - 2a_1 \exp(-a_2 \log p)$*,*

$$||\widehat{\beta} - \beta^*||_2 \le a_1^* \sqrt{s \log p / n} \quad \text{and} \quad ||\widehat{\beta} - \beta^*||_1 \le 4 a_1^* s \sqrt{\log p / n},$$

where $a_1^* = 4(2k_0 + c^*)/a^*$.

(ii) (Soft sparsity case) Assume β^* *satisfies the soft sparsity assumption (4). For any* R *satisfying* $R \geq \sqrt{\log p/n}$ *and* $c^{''}\sqrt{\log p/n}\max\{2, R\} < 1/2$, *where* $c^{''} = 4(2k_0 + c^*)/a^*$, *we have, with probability at least* $1 - 4\delta_n - 4\exp(-\log p) - 2a_1\exp(-a_2 \log p)$,

$$||\widehat{\beta} - \beta^*||_2 \leq a_2^* R^{1/2}(\log p/n)^{1/4} \quad \textit{and} \quad ||\widehat{\beta} - \beta^*||_1 \leq 4\big(a_2^*\sqrt{|S_a|}R^{1/2}(\log p/n)^{1/4} + ||\beta^*_{S_a^c}||_1\big),$$

where $a_2^* = 4\max\left\{2\sqrt{2}(2k_0+c^*)/a^*, \sqrt{(2k_0 + c^*)/a^*}\right\}$ *and* $|S_a|$ *denotes the cardinality of the set* $S_a = S_{-a} \cup \{1\}$ *with* $S_{-a} = \{j : |\beta_j^*| > a, 2 \leq j \leq p\}$, *and* a *being a positive thresholding parameter.*

Remark 3 (On the results of QR-LASSO for the hard sparsity case). In this case, the L_2 estimation error of QR-LASSO has the rate $\sqrt{s\log p/n}$. This matches the minimax optimal rate for LS-LASSO, established in Raskutti et al. (2011) under the assumption of sub-Gaussian errors, for the hard sparsity case. In the oracle case (when the underlying model is known), the L_2 estimation error has the rate $\sqrt{s/n}$. The above minimax rate is near-oracle up to a factor of order $\sqrt{\log p}$, the price for not knowing in advance which of the p covariates are relevant.

The results in Theorem 2(i) for the hard sparsity case are inspired by the earlier work of Belloni and Chernozhukov (2011) and Wang (2013), which obtained the same rates for the L_2 error bound. However, our proof employs a different technique and requires weaker conditions. Comparing with the conditions in Belloni and Chernozhukov (2011), we relaxed the conditions on both Σ and ϵ_i. We have dropped their *restricted nonlinearity condition* on Σ (their condition D.4), which would require $q := \inf_{\delta\in A, \delta\neq 0} \frac{\{E(|X_i^T\delta|^2)\}^{3/2}}{E(|X_i^T\delta|^3)} > 0$ for some restricted set A. Such a condition is not needed for the parallel theory of LS-LASSO. Furthermore, if the non-linear impact coefficient q converges to zero at a sufficiently fast rate, this may have a negative impact on the feasible range of n and p through the *growth condition* $\sqrt{s\log(p \vee n)} \leq O(q\sqrt{n})$ required in the main theorem (Theorem 2) of Belloni and Chernozhukov (2011). Unlike Belloni and Chernozhukov (2011), we do not require the conditional random error density $f_i(t)$ to be continuously differential nor the derivative to be uniformly bounded everywhere. We only need a uniform lower bound for $f_i(t)$ in a neighborhood of zero. Our assumptions are also significantly weaker than those in Wang (2013), which required independent and identically distributed random errors and a restricted isometry type condition in addition to the restricted eigenvalue condition.

Remark 4. (On the results of QR-LASSO for the soft sparsity case). The results in Theorem 2(ii) for the soft sparsity case are new for high-dimensional quantile regression. The soft sparsity scenario allows for dense small coefficients. The radius of the L_1-ball $\mathbb{B}_1(R)$ is allowed to shrink or diverge with the sample size n. In this case, we obtain the L_2 error rate $R^{1/2}(\log p/n)^{1/4}$ for QR-LASSO, which also matches the minimax optimal rate in the soft sparsity case for LS-LASSO, as shown in Raskutti et al. (2011). The L_1 error bound is larger than the L_2 error bound. However, one may still achieve an L_1 consistency rate under additional structural assumptions on β^*. For example, we consider an approximately sparse model. Without loss of generality, we assume $|\beta_2^*| \geq |\beta_3^*| \geq \ldots \geq |\beta_p^*|$, and that there exists a positive integer $q < p$ such that $q\sqrt{\log p/n} = o(1)$ and $\sum_{j=q+1}^{p} |\beta_j^*| = o(1)$. Then the result in Theorem 2(ii) implies that $||\widehat{\beta} - \beta^*||_1 = o(1)$. As suggested by an anonymous referee, we may consider fast decaying coefficients of the form $\beta_j^* = \theta^j$, for some $0 < \theta < 1$, $j = 2, \ldots, p$. Then taking $q = (n/\log p)^{1/3}$, we have $||\widehat{\beta} - \beta^*||_1 = o(1)$.

Remark 5. The regularization parameter λ is taken to be of the order $\sqrt{\log p/n}$, the universal penalty level introduced in Donoho and Johnstone (1994). The literature of regularized high-dimensional regression often focuses on statistical analysis with a penalty parameter of this order (e.g., Bickel et al. (2009)). In practice, an appealing approach (Belloni and Chernozhukov (2011)) is to directly simulate λ as the $(1-\alpha)$-quantile of the distribution of $c||\widetilde{S}_n||_\infty$, for some small $\alpha > 0$. This is feasible by observing that the distribution of $||\widetilde{S}_n||_\infty$ is pivotal. With this simulated choice of λ, the same estimation error bound would hold with probability at least $1 - \alpha - 4\delta_n - 4\exp(-\log p) - 2a_1\exp(-a_2\log p)$.

Remark 6. Lemma B.1 in Appendix B demonstrates that for a certain class of heavy-tailed error distributions and a general class of design matrices, there is a positive probability LS-LASSO cannot achieve the near-oracle rate derived under the normality assumption no matter the choice of the tuning parameter. In contrast, the results for regularized quantile regression in this paper hold for a much larger class of error distributions. Our results hence provide strong evidence for the robustness and broader applicability of quantile regression in high dimension.

Theorem 3 below gives the estimation error bounds for the feasible local solutions of QR-NCP. The nonconvex penalty function is assumed to satisfy some general conditions. The penalty function $p_\lambda(t)$ is defined on the real line and is symmetric about zero. It is assumed to be nondecreasing and concave for $t \in [0, +\infty)$, with a continuous derivative $p'_\lambda(t)$ on $(0, +\infty)$ and $\lim_{t\to 0+} p'_\lambda(t) = \lambda$. For $t > 0$, $p_\lambda(t)$ is nonincreasing in t. Furthermore, there exists a constant $\gamma_0 > 0$ such that the function $t \mapsto \rho_\lambda(t) + \frac{\gamma_0}{2}t^2$ is convex. This class of nonconvex penalty functions, in particular, include the popular choices SCAD and MCP penalties discussed in Section 2.1.

Theorem 3 *(QR-NCP) Let $\lambda = k_0\sqrt{\log p/n}$, where $k_0 \geq 4\max\{2\sqrt{m_x}, c^*\}$. Suppose conditions (C1)-(C3) are satisfied. Consider any feasible local solution $\widehat{\beta}$ such that $||\widehat{\beta}||_1 < \kappa$ for some $\kappa > ||\beta^*||_1$ and the KKT condition (7) is satisfied. Let a_1 and a_2 be the positive constants in Theorem 1.*

(i) (Hard-sparsity case) Assume β^* satisfies the hard sparsity assumption (3), with $\sqrt{\log p/n} < \frac{2a^*}{3\kappa k_0}$. Then with probability at least $1 - 4\delta_n - 4\exp(-\log p) - 2a_1\exp(-a_2\log p)$,

$$||\widehat{\beta} - \beta^*||_2 \leq a_3^*\sqrt{s\log p/n}, \quad \text{and} \quad ||\widehat{\beta} - \beta^*||_1 \leq 4a_3^* s\sqrt{\log p/n},$$

where $a_3^* = \frac{6k_0}{4a^* - 3\gamma_0}$.

(ii) (Soft sparsity case) Assume β^* satisfies the soft sparsity assumption (4). If $\sqrt{\log p/n} <$ $\max\{R, \frac{2a^*}{3\kappa k_0}\}$, then with probability at least $1 - 4\delta_n - 4\exp(-\log p) - 2a_1\exp(-a_2\log p)$,

$$||\widehat{\beta} - \beta^*||_2 \leq a_4^* R^{1/2}(\log p/n)^{1/4}, \quad \text{and} \quad ||\widehat{\beta} - \beta^*||_1 \leq 4a_4^*\sqrt{|S_a|}R^{1/2}(\log p/n)^{1/4} + 2||\beta^*_{S_a^c}||_1$$

where $a_4^* = 2\max\left\{\frac{3\sqrt{2}k_0}{2(a^* - \gamma_0)}, \sqrt{\frac{k_0}{a^* - \gamma_0}}\right\}$, and $|S_a|$ denotes the cardinality of index set S_a.

Remark 7. The assumption of the KKT condition (7) is satisfied for the SCAD and MCP penalty functions mentioned earlier. The side condition $||\widehat{\beta}||_1 \leq \kappa$ is imposed to focus on sensible local solutions as was adopted in Loh and Wainwright (2015). The high probability error bound in this theorem applies to any feasible local solution with the radius κ of the true value β^*. A careful examination of the proof reveals that κ can diverge to ∞ as long as $\kappa\sqrt{\log p/n} = o(1)$. It is also noted that we do not restrict the local solution to be sparse even for the hard sparsity setting. In practice, one would not wish to choose an exceedingly large κ, as the conditions of the theorem suggest that the non-asymptotic bounds hold with a larger n when κ is larger.

Embedded in the proof of Theorem 3 is a result on the geometric structure of the local solution of QR-NCP. Unlike the result in Lemma 2 about the geometric structure of the global solution of QR-Lasso, the result for QR-NCP in Lemma 1 below is new.

Corollary 1 *(QR-NCP) Assume the conditions of Theorem 3 are satisfied. Then with probability at least* $1 - 4\delta_n - 4\exp(-\log p) - 2a_1 \exp(-a_2 \log p)$*, where* a_1 *and* a_2 *are the positive constants in Theorem 1, we have (i)* $\widehat{v} \in \widetilde{\Gamma}_H$ *under the hard sparsity assumption; and (ii)* $\widehat{v} \in \widetilde{\Gamma}_W$ *under the soft sparsity assumption.*

4 Quantile prediction error bounds

Finally, we establish theoretical guarantees for the prediction error bounds of regularized quantile regression estimator in high dimension. Based on the check loss function, the empirical quantile prediction error is given by

$$R_n(\widehat{\beta}) = Q_n(\widehat{\beta}) - Q_n(\beta^*),$$

where $Q_n(\beta)$ is defined in (5). Our result does not impose restrictions on what regularized procedure is used. Specifically, the prediction error is evaluated using $\widehat{\beta}$, either the global solution from QR-LASSO or a local solution from QR-NCP.

A key result of this section is a general characterization of the error bound based on the subgradients. Consider any two subgradients $S_n(\beta), \overline{S}_n(\beta) \in \partial Q_n(\beta)$. The convexity of the loss function $Q_n(\beta)$ implies that

$$Q_n(\widehat{\beta}) \geq Q_n(\beta^*) + S_n(\beta^*)^T(\widehat{\beta} - \beta^*), \tag{17}$$

$$Q_n(\beta^*) \geq Q_n(\widehat{\beta}) + \overline{S}_n(\widehat{\beta})^T(\widehat{\beta} - \beta^*). \tag{18}$$

Combining inequalities (17) and (18), we immediately obtain an error bound based on subgradients:

$$|Q_n(\widehat{\beta}) - Q_n(\beta^*)| \leq \max\left\{||S_n(\beta^*)||_\infty, ||\overline{S}_n(\widehat{\beta})||_\infty\right\}||\widehat{\beta} - \beta^*||_1.$$

This leads to simple bounds of the prediction error for the general case (global or local solution, hard or soft sparsity).

Theorem 4 *Let $\lambda = k_0\sqrt{\log p/n}$, where $k_0 \geq 4\max\{2\sqrt{m_x}, c^*\}$. With probability at least $1 - \delta_n - 2\exp(-\log p)$, we have*

$$|R_n(\widehat{\beta})| \leq 4\lambda||\widehat{\beta} - \beta^*||_1. \tag{19}$$

Corollary 2 below summarizes the results for the quantile prediction error. These results are new for high-dimensional quantile regression. Our results extend the prediction error bound for the L_1-regularized least-squares regression, see Greenshtein et al. (2004), Bunea et al. (2007), Bickel et al. (2009) and Raskutti et al. (2011). Different from quantile regression, the precition error bound for LS-LASSO is usually studied based on the the least squares loss function $n^{-1}\sum_{i=1}^n(\epsilon_i - X_i^T\boldsymbol{\gamma})^2$. Our approach can also be applied to other (possibly nondifferentiable) convex loss functions.

Corollary 2

(i) (QR-LASSO) Let $\widehat{\beta}$ be the QR-Lasso estimator. Assume the conditions of Theorem 2 are satisfied. Then with probability at least $1 - 4\delta_n - 4\exp(-\log p) - 2a_1\exp(-a_2\log p)$,

$$|R_n(\widehat{\beta})| \leq b_1^* s\log p/n, \quad \textit{hard sparsity case,}$$

$$|R_n(\widehat{\beta})| \leq b_2^*\big(a_2^*\sqrt{|S_a|}R^{1/2}(\log p/n)^{3/4} + ||\beta^*_{S_a^c}||_1(\log p/n)^{1/2}\big), \quad \textit{soft sparsity case,}$$

where $b_1^* = 16k_0 a_1^*$, $b_2^* = 16k_0$, *and* S_a *is the same set as in Theorem 2.*

(ii) (QR-NCP) Let $\widehat{\beta}$ *be the QR-NCP estimator. Assume the conditions of Theorem 3 are satisfied. Then with probability at least* $1 - 4\delta_n - 4\exp(-\log p) - 2a_1 \exp(-a_2 \log p)$,

$$\begin{aligned} |R_n(\widehat{\beta})| &\leq b_3^* s \log p/n, \quad \textit{hard sparsity case,} \\ |R_n(\widehat{\beta})| &\leq b_4^* \big[2a_4^* \sqrt{|S_a|} R^{1/2} (\log p/n)^{3/4} + ||\beta^*_{S_a^c}||_1 (\log p/n)^{1/2}\big], \quad \textit{soft sparsity case,} \end{aligned}$$

where $b_3^* = 4k_0 a_3^*$ *and* $b_4^* = 8k_0$.

Remark 8. The rates for the hard sparsity case are the same as those in the literature for LS-LASSO. In the soft sparsity case, simplified upper bounds can be obtained by observing that $a|S_{-a}| \leq \sum_{i=2}^p |\beta_i^*| \leq R$, where $|S_{-a}|$ denotes the cardinality of the set S_{-a}. Hence $|S_{-a}| \leq a^{-1}R$. For QR-Lasso, since $\widehat{v} \in \Gamma_W$, we have $||\widehat{v}||_1 \leq 4\big(||\widehat{v}_{S_a}||_1 + ||\beta^*_{S_a^c}||_1\big) \leq 4\big(\sqrt{a^{-1}R + 1}||\widehat{v}||_2 + R\big)$, which holds for any $a > 0$. Taking $a = R/3$, we obtain $||\widehat{v}||_1 \leq 4\big[2||\widehat{v}||_2 + R\big]$. This also holds for QR-NCP. Hence,

$$\begin{aligned} |R_n(\widehat{\beta})| &\leq b_2^*[2a_2^* R^{1/2} (\log p/n)^{3/4} + R(\log p/n)^{1/2}], \quad \text{QR-Lasso,} \\ |R_n(\widehat{\beta})| &\leq 2b_4^*[4a_4^* R^{1/2} (\log p/n)^{3/4} + R(\log p/n)^{1/2}], \quad \text{QR-NCP.} \end{aligned}$$

The upper bounds for the soft sparsity case in Corollary 2 permits a faster rate when the true value β^* has certain desirable structural property: particularly when the number of relatively large signals in β^* is of a relatively small order, while the number of relatively small signals are much smaller relative to R. Finally, we note that for all the scenarios we considered, we require overall milder conditions on the random error distributions than those required in the literature for LS-LASSO.

5 A Monte Carlo experiment

In this section we carry out a Monte Carlo experiment to confirm some of the theoretical findings about LASSO or SCAD regularized quantile regression versus the least squares regression under the same form of penalization.

We first generate $(X_1, X_2, \cdots, X_p)$ from the multivariate normal distribution $N_p(0, \Sigma)$ with $\Sigma = (\sigma_{jk})_{p\times p}$ and $\sigma_{jk} = 0.5^{|j-k|}$, $1 \leq j, k \leq p$. For the regression parameter β^*, we consider two different models.

- Model 1 (sparser model): $\beta^* = (2, 1, 1.5, 1.75, 0_{p-4}^T)^T$,
- Model 2 (denser model): $\beta^* = \frac{3}{n}(1_n, 0_{p-n}^T)^T$,

where 0_k denotes the k-dimensional vector of zeros, while 1_k denotes the k-dimensional vector of ones. For each model, we consider two different random error distributions for ϵ_i: the $N(0, 1)$ distribution and the mixture normal distribution $aN(0, 1) + (1 - a)N(0, 10^2)$, where $a \sim$ Bernoulli(0.95).

We consider LS_Oracle, QR_Oracle, LS_LASSO, QR_LASSO, LS_SCAD (SCAD regularized least squares regression) and QR_SCAD (SCAD regularized quantile regression), where the quantile methods are based on $\tau = 0.5$, LS_Oracle and QR_Oracle are computed using the true model. For $n = 100$ and $p = 500$, the boxplots for the L_2-error of the six methods based on 1000 simulation runs for Model 1 and Model 2 are given in Figure 1 and Figure 2, respectively.

For the LASSO penalty, the tuning parameter was selected using a 5-fold cross-validation. For SCAD penalty, the tuning parameter is selected using a high-dimensional BIC procedure (Wang et al. (2013b), Lee et al. (2014), among others). We observe from Figure 1 that for the normal random error case LS-LASSO is slightly more efficient than QR-LASSO but its performance deteriorates substantially for the mixture normal random error case. The nonconvex SCAD penalty leads to smaller L_2 error than the LASSO penalty. Figure 2 suggests that for the denser model under consideration, QR-LASSO and QR-SCAD have similar performance to that of LS-LASSO and LS-SCAD for the normal error model and have much smaller error for the mixture normal error model.

Finally, Figure 3 depicts the L_2 errors for LS_LASSO, QR_LASSO, LS_SCAD and QR_SCAD for the two error distributions for Model 1 as the sample size varies between 100 and 800. The plot suggests the L_2 error decreases as n increases. It also suggests that for the heavy-tailed error distribution, regularized quantile regression substantially outperforms regularized least squares regression for all sample sizes.

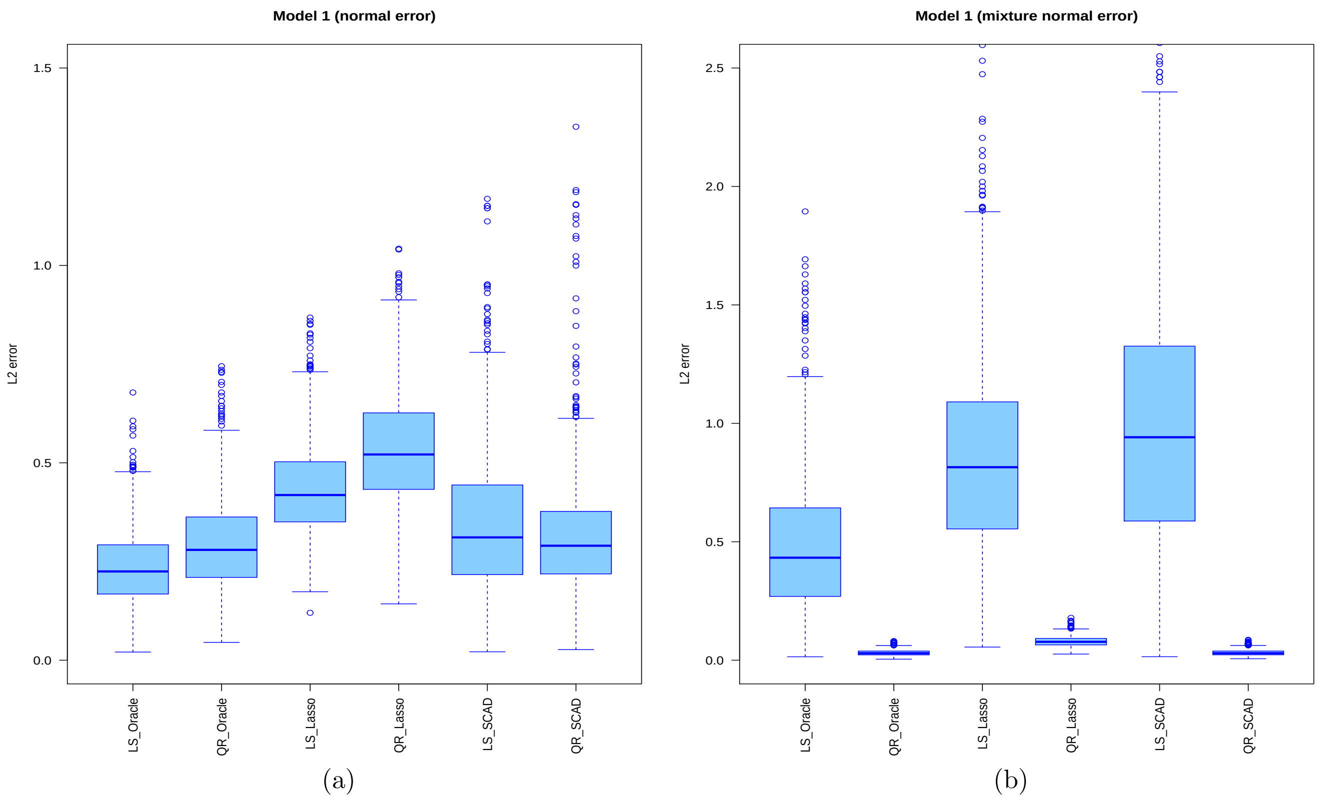


Figure 1: Box plots of the L_2 estimation error for Model 1 based on six methods: LS_Oracle, QR_Oracle, LS_LASSO, QR_LASSO, LS_SCAD and QR_SCAD.

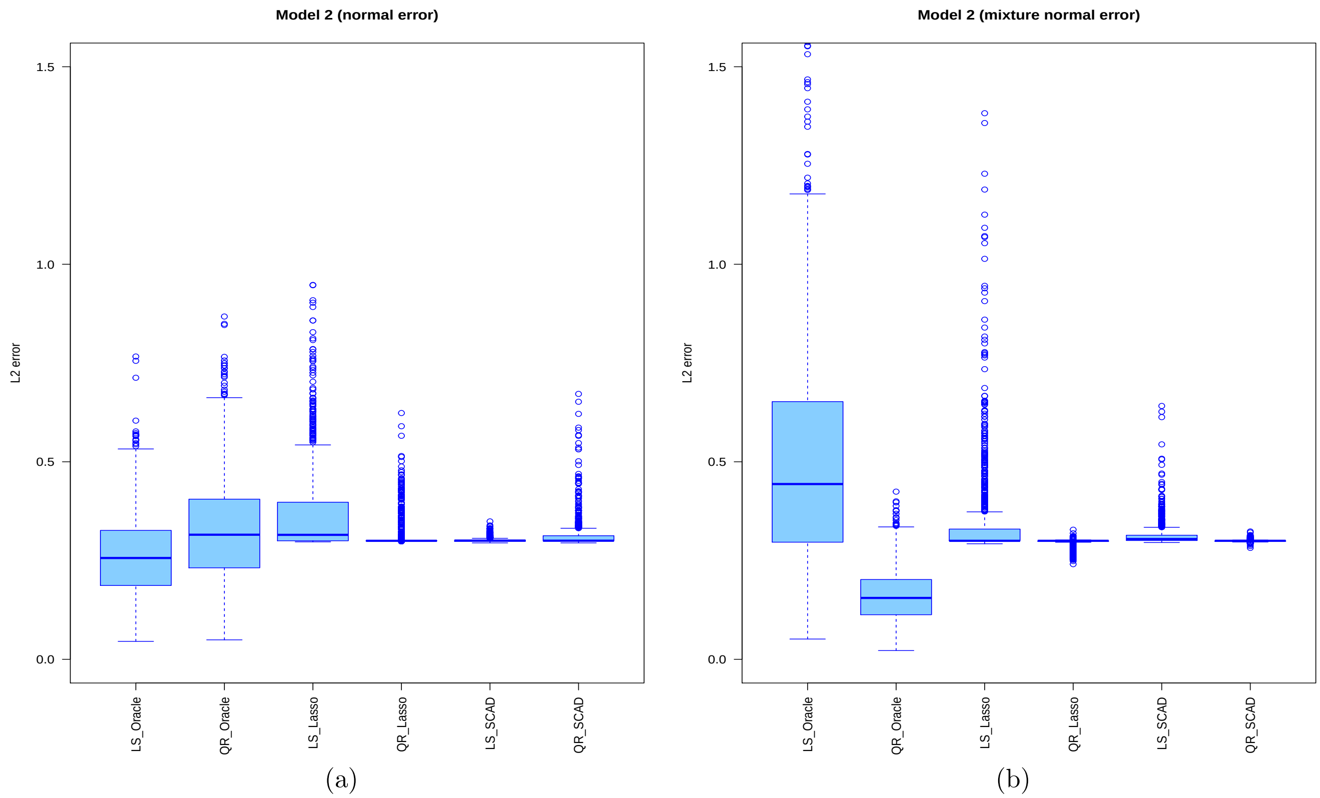
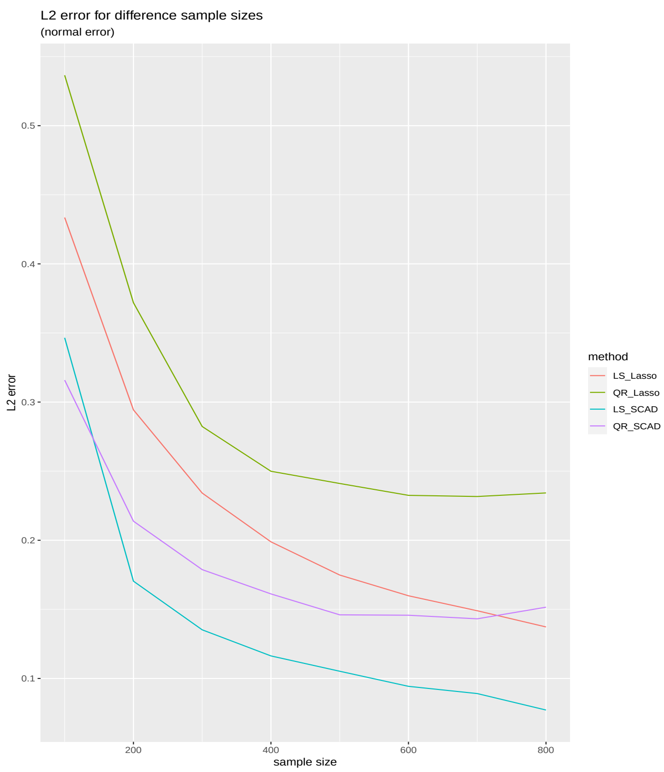
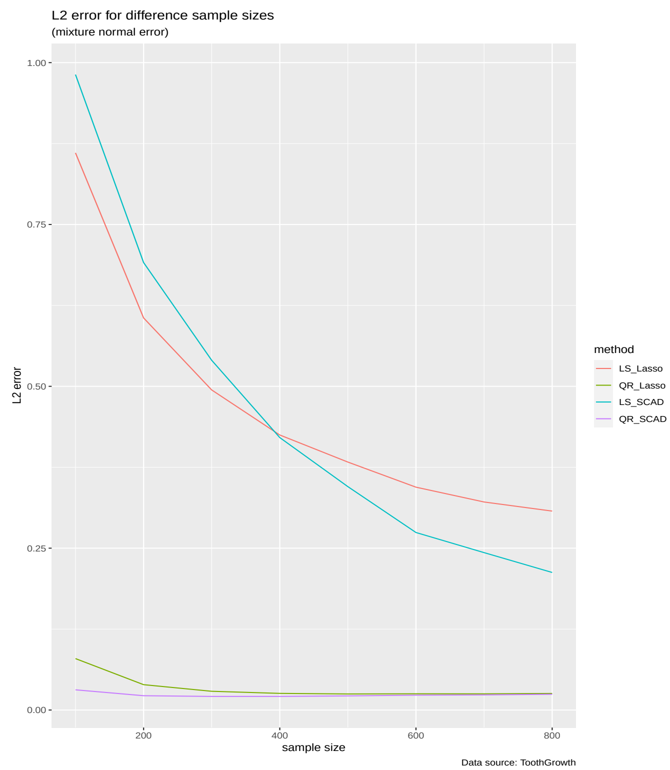


Figure 2: Box plots of the L_2 estimation error for Model 2 based on six methods: LS_Oracle, QR_Oracle, LS_LASSO, QR_LASSO, LS_SCAD and QR_SCAD.



(a)



(b)

Figure 3: The L_2 estimation error for Model 1 with varying sample sizes.

6 Discussions

By developing a new and unified subgradient approach, we present several significant results on the fundamental properties of regularized quantile regression in high dimension, where the number of covariates can grow at an exponential rate of the sample size. We demonstrate that L_1 regularized quantile regression enjoys the near-oracle rate in estimation error for a much richer class of error distributions than does LASSO (with least squares). The result renders theoretical support for the wide applicability of quantile regression in high dimensional problems. Our analysis is carried out under both the hard sparsity and the soft sparsity models. We also prove that any feasible local solution of quantile regression with some commonly used nonconvex penalties such as SCAD and MCP enjoys the same estimation error rates, making the theoretical results more meaningful to practical solutions that might reach a local optimum instead of a global one.

The unified approach based on subgradients can be useful for studying other high-dimensional problems involving a nonsmooth loss function, so it is of independent interest. For example, the support vector machine (SVM) is a powerful binary classification tool with high accuracy and great flexibility. The sample loss function for SVM is given by

$$n^{-1}\sum_{i=1}^{n}(1 - Y_i\mathbf{X}_i^T\boldsymbol{\beta})_+$$

where $Y_i \in \{1, -1\}$, and $(1-u)_+ = \max\{1-u, 0\}$ denotes the hinge loss. As another example, we consider the rank loss function for robust high-dimensional regression

$$\big[n(n-1)\big]^{-1}\sum\sum_{i\neq j}\big|(Y_i - \mathbf{x}_i^T\beta) - (Y_j - \mathbf{x}_j^T\beta)\big|.$$

Both examples involve non-smooth loss functions. The sub-gradient techniques developed in this paper can be applied to study the statistical properties of the global and local solutions of the corresponding regularized estimation problems in high dimension when $p \gg n$.

The results of the paper are also useful for inference procedures for high-dimensional quantile regression, which need an initial value, see for example Belloni et al. (2014), Zhao et al. (2014), Belloni et al. (2019).

Appendix A Identification of parameters

Let $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_p)$ be the $n \times p$ matrix of covariates. In the classical asymptotic framework where p is smaller than n, it is usually assumed that $\mathbb{X}$ has a full column rank, which allows to uniquely identify β^*. In contrast, in the high-dimensional scenario where $p \gg n$, β^* is generally not identifiable in the absence of additional structural assumption as $\mathbb{X}$ has at most column rank n.

In high dimension, β^* in general is not uniquely defined. Suppose model (1) is satisfied by $\beta^* = \beta_0^*$. Consider the affine space $\{\beta^* \in \mathbb{R}^p : \mathbb{X}\beta^* = \mathbb{X}\beta_0^*\}$. *We emphasize that* the error bounds derived in this paper apply to any β^* from this affine space and does not require the unique identification.

Let $\text{Ker}(\mathbb{X}) = \{\beta \in \mathbb{R}^p : \mathbb{X}\beta = \mathbf{0}\}$ be the null space of $\mathbb{X}$. If β^* satisfies (1), then $\beta^* + \beta$ also satisfies (1), $\forall\ \beta \in \text{Ker}(\mathbb{X})$. The extent of identifiability can be measured by the diameter of the set $N_0(\mathbb{X}) = \text{Ker}(\mathbb{X}) \cap \mathbb{B}$, defined as $\max_{\beta \in N_0(\mathbb{X})} ||\beta||_2$, where $\mathbb{B} = \mathbb{B}_0(s)$ under the hard sparsity assumption, while $\mathbb{B} = \mathbb{B}_1(R)$ under the soft sparsity assumption; The following lemma characterizes the properties of the diameter of $N_0(\mathbb{X})$.

Lemma A.1 *Assume the vector $X_{-i} = (X_{i2}, \ldots, X_{ip})'$ is a mean-zero sub-Gaussian random vector.*

(i) (hard sparsity case) Assume $\eta_{min}(s) = \inf_{v:||v||_2=1,||v||_0\le s} v'\Sigma v > 0$, where $\eta_{min}(s)$ is the smallest s-sparse eigenvalue of $\Sigma = E(X_i X_i')$. Then

$$P\Big(\max_{\beta \in N_0(\mathbb{X})} ||\beta||_2 = 0\Big) \ge 1 - \alpha_1^* \exp(-\alpha_2^* \log p),$$

where α_1^* and α_2^* are positive constants.

(ii) (soft sparsity case),

$$P\Big(\max_{\beta \in N_0(\mathbb{X})} ||\beta||_2 \le \alpha' \xi_{min}^{-1/2} R \sqrt{\log p/n}\Big) \ge 1 - \alpha_1^* \exp(-\alpha_2^* \log p),$$

where α_1^*, α_2^* and α' are positive constants, and ξ_{min} is the smallest eigenvalue of Σ.

The above lemma can be considered as a generalization of Lemma 1 of Raskutti et al.

(2011) to the random covariates case. From this lemma, it can be seen that under some general conditions, $\mathrm{Ker}(\mathbb{X}) = \{\mathbf{0}\}$ with high probability for the hard sparsity case, i.e., the sparse β^* satisfying (1) is unique. While for the soft sparsity case, $\mathrm{Ker}(\mathbb{X})$ is a shrinking neighborhood around $\mathbf{0}$ with high probability if $\xi_{min}^{-1/2} R\sqrt{\log p/n} \to 0$.

Proof of Lemma A.1 (i) (hard sparsity case) In Lemma D.1, take $k = s$ and consider $\mathbb{C}_1(s) = \{\theta \in \mathbb{R}^p : ||\theta||_2 = 1, ||\theta||_0 \leq s\}$. Also in this lemma, take $t = \eta_{min}(s)/2$, where $\eta_{min}(s)$ denotes the smallest s-sparse eigenvalue of Σ, i.e., $\eta_{min}(s) = \inf_{v:||v||_2=1,||v||_0\leq s} v'\Sigma v$. We have

$$\begin{aligned} & P\Big\{ \sup_{\theta\in\mathbb{C}_1(s)} \big| n^{-1}||\mathbb{X}\theta||_2^2 - E(n^{-1}||\mathbb{X}\theta||_2^2)\big| \geq \eta_{min}(s)/2 \Big\} \\ \leq \;& \alpha_2 \exp\Big(- \alpha_1 n \min\big(\eta_{min}^2(s)/(4\sigma_x^4), \eta_{min}(s)/(2\sigma_x^2)\big) + 2s\log p \Big), \end{aligned} \tag{A.1}$$

for some positive constants α_1 and α_2. We argue by contradiction and assume $\max_{\beta\in N_0(\mathbb{X})} ||\beta||_2 \neq 0$. Then there exists a $\theta \neq 0$ in $\mathbb{B}_0(s)$ such that $\mathbb{X}\theta = 0$. Let $\tilde{\theta} = \theta/||\theta||_2$, then $\tilde{\theta} \in \mathbb{C}_1(s)$. It follows from (A.1), there exist some positive constants α_1^* and α_2^* such that with probability at least $1 - \alpha_1^* \exp(-\alpha_2^* n)$,

$$n^{-1}||\mathbb{X}\tilde{\theta}||_2^2 \geq \tilde{\theta}'\Sigma\tilde{\theta} - \eta_{min}(s)/2 \geq \eta_{min}(s)/2 > 0,$$

which contradicts with the assumption $\mathbb{X}\tilde{\theta} = 0$.
(ii) (soft sparsity case) In Lemma D.1, we take $t = \xi_{min}/54$ and $k = \frac{1}{4}\alpha_1 \min(t^2/\sigma_x^4, t/\sigma_x^2)\frac{n}{\log p}$, where ξ_{min} denotes the smallest eigenvalue of Σ. We have

$$P\Big\{ \sup_{\theta\in\mathbb{C}(k)} \big| n^{-1}||\mathbb{X}\theta||_2^2 - E(n^{-1}||\mathbb{X}\theta||_2^2)\big| \geq t \Big\} \;\; \leq \;\; \alpha_2 \exp\big(- \alpha_1 n \min(t^2/\sigma_x^4, t/\sigma_x^2)/2 \big),$$

for some positive constants α_1 and α_2. Then the same argument as in Lemma 11 and Lemma 13 of Loh and Wainwright (2012) leads to

$$n^{-1}||\mathbb{X}\theta||_2^2 \geq \frac{\xi_{min}}{2}||\theta||_2^2 - c'n^{-1}(\log p)||\theta||_1^2,$$

for all $\theta \in \mathbb{R}^p$, with probability at least $1 - \alpha_1^* \exp\big(-\alpha_2^* n\big)$, for some positive constants c', α_1^* and α_2^*. This implies that for any $\theta \in \mathrm{Ker}(\mathbb{X}) \cap \mathbb{B}(R)$,

$$0 = n^{-1}||\mathbb{X}\theta||_2^2 \geq \frac{\xi_{min}}{2}||\theta||_2^2 - c'R^2 n^{-1}\log p,$$

or $\max_{\theta\in N_0(\mathbb{X})}||\theta||_2 \leq \alpha'\xi_{min}^{-1/2}R\sqrt{\log p/n}$ for some positive constant α' with probability at least $1 - \alpha_1^* \exp\big(-\alpha_2^* n\big)$. Hence the conclusion follows. □

Appendix B Performance lower bound for Lasso with heavy-tailed errors

The LS-LASSO estimator $\widehat{\beta}^{LS}$ is defined as

$$\widehat{\beta}^{LS} = \arg\min_{\beta\in\mathbb{R}^p}\Big\{(2n)^{-1}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda||\beta_-||_1\Big\}. \tag{B.1}$$

Suppose $\epsilon_i = Y_i - X_i^T\beta^*$ $(i = 1, \cdots, n)$ are independent Cauchy(0,1) random variables with the density function $f(\epsilon) = \frac{1}{\pi(1+\epsilon^2)}$, and the characteristic function $\phi(t) = \mathrm{E}[e^{i\epsilon t}] = e^{-|t|}$, $\forall\ t \in \mathbb{R}$. The covariates $X_i = (x_{i1}, \ldots, x_{ip})^T$ are independent p-dimensional sub-Gaussian random vectors with variance proxy ζ_0^2, that is, $\forall\ v \in \mathbb{R}^p$, $\forall\ t \in \mathbb{R}$, $\mathrm{E}\{\exp(tX_i^Tv)\} \leq \exp\{\zeta_0^2t^2||v||_2^2/2\}$, where ζ_0 is a positive constant. Assume $\min_{1\leq j\leq p}\mathrm{E}(|x_{ij}|) \geq \zeta_* > 0$. The $\widehat{\beta}^{LS}$ estimator is called a non-degenerate solution if it has at least one non-zero component. We consider the asymptotic regime where $\log p = o(n)$.

Lemma B.1 *Consider the setting described above. Let a_0 be any positive constant and consider an arbitrary $\lambda \in (0, a_0)$. Let $\widehat{\beta}^{LS}$ be any non-degenerate solution of (B.1) corresponding to λ. Then there exist some constants $\zeta_1 > 0$ and $0 < \zeta_2 < 1$ such that if (n, p) satisfies $p \geq 5$, $\log p \leq n/4$ and $\sqrt{\log p/n} \leq \zeta_1$, then*

$$P\big(||\widehat{\beta}^{LS} - \beta^*||_1 > 1\big) \geq \zeta_2,$$

where the constants ζ_1 and ζ_2 do not depend on (n, p).

Loh (2017) showed that if one chooses the tuning parameter λ at the regular rate $\sqrt{\log p/n}$, which is the theoretical choice that leads to the near-oracle performance of LS-LASSO for normal random errors, then the KKT condition for LS-LASSO may not hold for Cauchy random errors. The above lemma strengthens the result by showing that a different choice of λ can not fix the problem. Indeed, for any $\lambda \in (0, a_0)$, the L_1-estimation error of LS-LASSO has a positive probability to exceed one. This paper shows that QR-Lasso is still consistent in this case with $\lambda \sim \sqrt{\log p/n}$. This lemma hence renders strong support for the robustness of QR-LASSO in high dimension. Related to this, Fan et al. (2014) showed that for a specially designed fixed design matrix, LS-LASSO cannot be sign-consistent unless a certain signal condition is satisfied.

It is worth noting that the above probability bound is nonasymptotic and holds for all n sufficiently large and does not become smaller as n increases. Although this lemma focuses on the Cauchy random error for a clean presentation, analogous inconsistency result holds more generally for the class of α-stable distributions with $\alpha \in (0, 2)$. Specifically, ϵ_i has an α-stable distribution with scale parameter ξ if its characteristic function $E\{\exp(it\epsilon_i)\} = \exp(-\xi^\alpha |t|^\alpha),\ \forall\ t \in \mathbb{R}^p$. The standard Cauchy distribution is an α-stable distribution with $\alpha = \xi = 1$, see Nolan (2003) for an introduction to stable distributions.

Proof of Lemma B.1. Assume the contrary is true, that is, $||\widehat{\beta}^{LS} - \beta^*||_1 \leq 1$ for some $\lambda \in (0, a_0)$. As $\widehat{\beta}^{LS}$ is a non-degenerate solution, it has at least one non-zero component. Without loss of generality, we assume $\widehat{\beta}_j^{LS} \neq 0$ for some $1 \leq j \leq p$. By the Karush–Kuhn–Tucker(KKT) condition, $\widehat{\beta}^{LS}$ must satisfy

$$e_j^T(n^{-1}\mathbb{X}^T\mathbb{X})(\beta^* - \widehat{\beta}^{LS}) + e_j^T n^{-1}\mathbb{X}^T\epsilon + \lambda \mathrm{sign}(\widehat{\beta}_j^{LS}) = 0, \tag{B.2}$$

where e_j is a p-dimensional unit vector with the j entry being one and all the other entries being zero, $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_p)$, is the $n \times p$ matrix of covariates.

Consider the event $\Omega_{n1} = \{||\widehat{\Sigma}||_\infty \leq 12\zeta_0^2\}$. By Lemma D.2, $P(\Omega_{n1}) \geq 1 - 2\exp(-\log p)$.

On Ω_{n1},

$$\begin{aligned}|e_j^T(n^{-1}\mathbb{X}'\mathbb{X})(\beta^*-\widehat{\beta}^{LS})| &\le ||e_j^T(n^{-1}\mathbb{X}'\mathbb{X})||_\infty||(\beta^*-\widehat{\beta}^{LS})||_1\\ &\le ||\widehat{\Sigma}||_\infty||(\beta^*-\widehat{\beta}^{LS})||_1\\ &\le 12\zeta_0^2||(\beta^*-\widehat{\beta}^{LS})||_1.\end{aligned}$$

As $|\lambda\text{sign}(\widehat{\beta}_j^{LS})|\le a_0$, it follows from the KKT condition (B.2) that on the event Ω_{n1} we have $|e_j^T n^{-1}\mathbb{X}^T\epsilon|\le a_0+12\zeta_0^2$.

Write $\mathbb{X}=(\tilde{X}_1,\ldots,\tilde{X}_p)^T$. Conditional on $\tilde{X}_j$, $e_j^T n^{-1}\mathbb{X}^T\epsilon = n^{-1}\tilde{X}_j^T\epsilon$ has a Cauchy(0, $n^{-1}\sum_{i=1}^n|x_{ij}|$) distribution, by checking the form of its characteristic function. By the property of Cauchy distribution, we have

$$P\big(|e_j^T n^{-1}\mathbb{X}^T\epsilon|>a_0+12\zeta_0\big)=2E_{\tilde{X}_j}\Big\{\frac{1}{2}-\frac{1}{\pi}\arctan\Big(\frac{a_0+12\zeta_0}{n^{-1}\sum_{i=1}^n|x_{ij}|}\Big)\Big\}.$$

Note that $n^{-1}\sum_{i=1}^n|x_{ij}|$ on Ω_n, $e_j^T n^{-1}\mathbb{X}^T\epsilon$ has a Cauchy(0, b) distribution with $0<b<12\zeta_0$. Consider the event $\Omega_{n2}=\big\{\min_{1\le j\le p} n^{-1}\sum_{i=1}^n|x_{ij}|\ge\zeta_*/2\big\}$. By Lemma D.3, $P(\Omega_{n2})\ge 1-2\exp(-\log p)$. On the event Ω_{n2},

$$P\big(|e_j^T n^{-1}\mathbb{X}'\epsilon|>a_0+12\zeta_0\big)\ge 1-\frac{2}{\pi}\arctan\big(\zeta_*^{-1}(2a_0+24\zeta_0)\big).$$

Hence on the event $\Omega_{n1}\cap\Omega_{n2}$, with probability at least $1-\frac{2}{\pi}\arctan\big(\zeta_*^{-1}(2a_0+24\zeta_0)\big)$, there is a contradiction. Hence,

$$\begin{aligned}P\big(||\beta^*-\widehat{\beta}^{LS}||_1>1\big) &\ge P\big(||\beta^*-\widehat{\beta}^{LS}||_1>1\big|\Omega_{n1}\cap\Omega_{n2}\big)P(\Omega_{n1}\cap\Omega_{n2})\\ &\ge \big(1-\frac{2}{\pi}\arctan\big(\zeta_*^{-1}(2a_0+24\zeta_0)\big)\big)(1-4\exp(-\log p))\\ &\ge \big(1-\frac{2}{\pi}\arctan\big(\zeta_*^{-1}(2a_0+24\zeta_0)\big)\big)/5.\end{aligned}$$

The result of the lemma holds with the ζ_1 defined in Lemma D.3 and $\zeta_2=\big(1-\frac{2}{\pi}\arctan\big(\zeta_*^{-1}(2a_0+24\zeta_0)\big)\big)/5$. $\square$

Appendix C Proof of the main theory

We provide below proofs of Theorems 1-4. Proofs of other results are given in Appendix D.

To prove Theorem 1, we will first establish that the lower bound stated in the theorem holds with high probability for

$$U_n(\Delta) = n^{-1}\sum_{i=1}^{n} X_i^T\Delta[\mathrm{I}(\epsilon_i \le X_i^T\Delta) - \mathrm{I}(\epsilon_i \le 0)], \tag{C.1}$$

which is $S_n(\beta^* + \Delta) - S_n(\beta^*)$ corresponding to a specific choice of subgradient in $\partial Q_n(\beta)$. Note that $U_n(\Delta)$ is not Lipschitz continuous in Δ. We start by constructing a Lipschitz continuous lower bound of $U_n(\Delta)$.

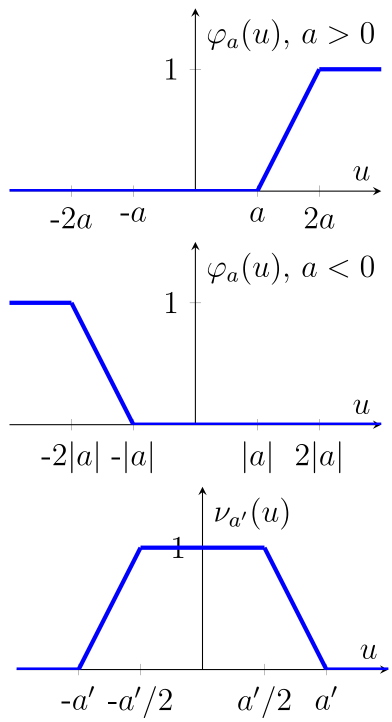
We observe that $X_i^T\Delta$ and $\mathrm{I}(\epsilon_i \le X_i^T\Delta) - \mathrm{I}(\epsilon_i \le 0)$ always have the same sign. Furthermore, if $X_i^T\Delta > 0$, $\mathrm{I}(\epsilon_i \le X_i^T\Delta) - \mathrm{I}(\epsilon_i \le 0)$ is nonzero only $0 < \epsilon_i \le X_i^T\Delta$; similarly if $X_i^T\Delta < 0$, $\mathrm{I}(\epsilon_i \le X_i^T\Delta) - \mathrm{I}(\epsilon_i \le 0)$ is nonzero only if $X_i^T\Delta < \epsilon_i \le 0$. Hence,

$$\begin{aligned}
U_n(\Delta) &= n^{-1}\sum_{i=1}^{n} X_i^T\Delta\mathrm{I}(0 < \epsilon_i \le X_i^T\Delta) - n^{-1}\sum_{i=1}^{n} X_i^T\Delta\mathrm{I}(X_i^T\Delta < \epsilon_i \le 0) \\
&\ge n^{-1}\sum_{i=1}^{n} |\epsilon_i|\big[\mathrm{I}(0 < \epsilon_i < X_i^T\Delta) + \mathrm{I}(X_i^T\Delta < \epsilon_i \le 0)\big] \\
&\ge n^{-1}\sum_{i=1}^{n} |\epsilon_i|\varphi_{\epsilon_i}(X_i^T\Delta)\nu_{b||\Delta||_2}(X_i^T\Delta),
\end{aligned}$$

where b is a positive constant, and

$$\varphi_a(u) = \begin{cases} 1, & \text{if } |u| > 2|a| \text{ and } ua > 0 \\ -1 + \frac{|u|}{a}, & \text{if } |a| \le |u| \le 2|a| \text{ and } ua > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\nu_{a'}(u) = \begin{cases} 1, & \text{if } |u| < \frac{a'}{2} \\ 2 - \frac{2|u|}{a'}, & \text{if } 0 < \frac{a'}{2} \le |u| \le a' \\ 0, & \text{otherwise} \end{cases}$$



for some $a \in \mathbb{R}$ and $a' \geq 0$. Note that $a\varphi_a(u)$ and $a'\nu_{a'}(u)$ are both Lipschitz in u: $|a\varphi_a(u_1) - a\varphi_a(u_1)| \leq |u_1 - u_2|$ and $|a'\nu_{a'}(u_1) - a'\nu_{a'}(u_1)| \leq 2|u_1 - u_2|$. We also note that

$$\mathrm{I}(u > 2a > 0) + \mathrm{I}(u < 2a < 0) \leq \varphi_a(u) \leq \mathrm{I}(|u| > |a|), \tag{C.2}$$

$$\nu_{a'}(u) \leq \mathrm{I}(|u| \leq a'), \tag{C.3}$$

$$1 - \nu_{a'}(u) \leq \mathrm{I}(|u| \geq a'/2). \tag{C.4}$$

Define

$$V_{nb}(\Delta) = n^{-1} \sum_{i=1}^{n} |\epsilon_i| \varphi_{\epsilon_i}(X_i^T \Delta) \nu_{b||\Delta||_2}(X_i^T \Delta).$$

Then $U_n(\Delta) \geq V_{nb}(\Delta)$, $\forall\ b > 0$. As Lemma C.1 below suggests, $V_{nb}(\Delta)$ is a Lipschitz-continuous lower bound of $U_n(\Delta)$.

Lemma C.1 *Let $g_\epsilon(\xi) = |\epsilon| \varphi_\epsilon(\xi) \nu_{b\delta}(\xi)$, where $\delta \geq 0$ is a constant. Then $g_\epsilon(\xi)$ is Lipschitz continuous in ξ,*

$$\left| g_\epsilon(\xi) - g_\epsilon(\xi') \right| \leq 3|\xi - \xi'|. \tag{C.5}$$

Proof. Case I: When $\delta = 0$, we have $\nu_{b\delta}(\xi) = \nu_{b\delta}(\xi') = 0$, and (C.5) is satisfied. So we only need to consider the scenario where $\delta > 0$.

Case II: $|\xi| > b\delta > 0$ and $|\xi'| > b\delta > 0$. In this case, $\nu_{b\delta}(\xi) = \nu_{b\delta}(\xi') = 0$. Hence, (C.5) holds trivially.

Case III: Assume $\delta > 0$ but case II is not satisfied. Then at least one of ξ or ξ' has absolute value no larger than $b\delta$. Without loss of generality, we assume $|\xi| \leq b\delta$. Then

$$\begin{aligned}
\big|g_\epsilon(\xi) - g_\epsilon(\xi')\big| &\leq \underbrace{|\epsilon|\varphi_\epsilon(\xi)}_{\leq\ |\xi|\ \leq\ b\delta}\ \underbrace{\big|\nu_{b\delta}(\xi) - \nu_{b\delta}(\xi')\big|}_{\leq\ \frac{2}{b\delta}|\xi-\xi'|} + \underbrace{|\epsilon|\big|\varphi_\epsilon(\xi) - \varphi_\epsilon(\xi')\big|}_{\leq\ |\xi-\xi'|}\ \underbrace{\nu_{b\delta}(\xi')}_{\leq 1} \\
&\leq 3|\xi - \xi'|.
\end{aligned}$$

Hence (C.5) holds. On the other hand, if $|\xi'| \leq \tau\delta$, we can bound $|g_\epsilon(\xi) - g_\epsilon(\xi')|$ by $|\epsilon||\varphi_\epsilon(\xi) - \varphi_\epsilon(\xi')|\ \nu_{b\delta}(\xi) + |\epsilon|\varphi_\epsilon(\xi')|\nu_{b\delta}(\xi) - \nu_{b\delta}(\xi')|$ and the conclusion follows similarly. $\square$

Lemma C.2 *Assume conditions (C1) and (C2) hold, where the positive constant b_0 in (C1) satisfies (C.7) below. Then there exists a positive constant k_0 such that*

$$E(V_{nb}(\Delta)) \geq a^*||\Delta||_2^2,$$

uniformly on $\{||\Delta||_2 \leq 1\} \cap \mathbb{C}$, $\forall\ b \geq b_0$.

Proof. We will first prove, by an application of the dominated convergence theorem, that there exists a positive constant b_0 such that $\forall\ b \geq b_0$,

$$\mathrm{E}_{X_i}\Big\{(X_i^T\Delta)^2\mathrm{I}\big(|X_i^T\Delta| \geq b||\Delta||_2/2\big)\Big\} \leq \frac{1}{2}\mathrm{E}\big\{(X_i^T\Delta)^2\big\}. \tag{C.6}$$

Write

$$\begin{aligned}
&\mathrm{E}_{X_i}\Big\{(X_i^T\Delta)^2\mathrm{I}\big(|X_i^T\Delta| \geq b||\Delta||_2/2\big)\Big\} \\
=\ &\mathrm{E}\big\{(X_i^T\Delta)^2\big\}\mathrm{E}_{X_i}\Big\{\frac{(X_i^T\Delta)^2}{\mathrm{E}\big\{(X_i^T\Delta)^2\big\}}\mathrm{I}\big(|X_i^T\Delta| \geq b||\Delta||_2/2\big)\Big\}.
\end{aligned}$$

We first note that $\mathrm{E}\Big\{\frac{(X_i^T\Delta)^2}{\mathrm{E}\big\{(X_i^T\Delta)^2\big\}}\Big\} \le 1$. Furthermore,

$$\begin{aligned}
& P\Big\{\frac{(X_i^T\Delta)^2}{\mathrm{E}\big\{(X_i^T\Delta)^2\big\}}\mathrm{I}\big(|X_i^T\Delta| \ge b||\Delta||_2/2\big) \ne 0\Big\} \\
\le\ & P\big\{|X_i^T\Delta| \ge b||\Delta||_2/2\big\} \\
\le\ & \frac{4\mathrm{E}\big\{(X_i^T\Delta)^2\big\}}{b^2||\Delta||_2^2} \le \frac{4k_u}{b^2}.
\end{aligned}$$

By the dominated convergence theorem, $\mathrm{E}_{X_i}\Big\{\frac{(X_i^T\Delta)^2}{\mathrm{E}\big\{(X_i^T\Delta)^2\big\}}\mathrm{I}\big(|X_i^T\Delta| \ge \frac{b||\Delta||_2}{2}\big)\Big\} \to 0$ as $b \to \infty$. Hence there exists a positive constant b_0, whose choice only depends on the probability distributions of X_i, such that $\forall\ b \ge b_0$, we have

$$\mathrm{E}_{X_i}\Big\{\frac{(X_i^T\Delta)^2}{\mathrm{E}\big\{(X_i^T\Delta)^2\big\}}\mathrm{I}\Big(|X_i^T\Delta| \ge \frac{b||\Delta||_2}{2}\Big)\Big\} \le 1/2. \tag{C.7}$$

We have

$$\begin{aligned}
& \mathrm{E}\big(V_{nb}(\Delta)\big) \\
=\ & n^{-1}\sum_{i=1}^n \mathrm{E}\big\{|\epsilon_i|\varphi_{\epsilon_i}(X_i^T\Delta)\nu_{b||\Delta||_2}(X_i^T\Delta)\big\} \\
\ge\ & n^{-1}\sum_{i=1}^n \mathrm{E}_{X_i}\Big\{\mathrm{E}_{\epsilon_i|X_i}\big[|\epsilon_i|\ \big\{\mathrm{I}\big(X_i^T\Delta > 2\epsilon_i > 0\big) + \mathrm{I}\big(X_i^T\Delta < 2\epsilon_i < 0\big)\big\}\big]\mathrm{I}\big(|X_i^T\Delta| \le b||\Delta||_2/2\big)\Big\} \\
=\ & n^{-1}\sum_{i=1}^n \mathrm{E}_{X_i}\Big\{\Big(\int_0^{|X_i^T\Delta|/2} tf_i(t)dt + \int_{-|X_i^T\Delta|/2}^0 (-t)f_i(t)dt\Big)\mathrm{I}\big(|X_i^T\Delta| \le b||\Delta||_2/2\big)\Big\} \\
=\ & n^{-1}\sum_{i=1}^n \mathrm{E}_{X_i}\Big\{\Big(\int_0^{|X_i^T\Delta|/2} t\big[f_i(t)dt + f_i(-t)\big]dt\Big)\mathrm{I}\big(|X_i^T\Delta| \le b||\Delta||_2/2\big)\Big\} \\
\ge\ & \frac{m_0}{4}\mathrm{E}_{X_i}\Big\{(X_i^T\Delta)^2\mathrm{I}\big(|X_i^T\Delta| \le b||\Delta||_2/2\big)\Big\} \\
=\ & \frac{m_0}{4}\mathrm{E}_{X_i}\Big\{(X_i^T\Delta)^2\Big\} - \frac{m_0}{4}\mathrm{E}_{X_i}\Big\{(X_i^T\Delta)^2\mathrm{I}\big(|X_i^T\Delta| > b||\Delta||_2/2\big)\Big\} \\
\ge\ & \frac{m_0}{8}\mathrm{E}_{X_i}\Big\{(X_i^T\Delta)^2\Big\} \\
\ge\ & a^*||\Delta||_2^2,
\end{aligned}$$

where $a^* = (m_0m_1)/8$, the first inequality applies (C.2) and (C.4), the second inequality applies condition (C1), the second last inequality applies (C.6) and the last inequality follows from condition (C2). $\square$

For an arbitrary $0 < \delta \leq 1$, let $S_2(\delta) = \{\Delta : ||\Delta||_2 = \delta\}$. Let $\Gamma(t) = \{\Delta : \Delta \in S_2(\delta),\ ||\Delta||_1 \leq t||\Delta||_2\} \cap \mathbb{C}$, for an arbitrary $t > 0$. Lemma C.3 below establishes a concentration inequality for

$$V_{nb}(\Delta) = n^{-1} \sum_{i=1}^{n} |\epsilon_i| \varphi_{\epsilon_i}(X_i^T \Delta) \nu_{b\,||\Delta||_2}(X_i^T \Delta),$$

where $b \geq b_0$ is a positive constant. Consider the event $A_{n1} = \{\max_{1\leq j \leq p} \widehat{\sigma}_j^2 \leq m_x\}$, where $\widehat{\sigma}_j^2 = n^{-1} \sum_{i=1}^{n} x_{ij}^2$. Then $P(A_{n1}) \geq 1 - \delta_n$, where $\delta_n \to 0$ as $n \to \infty$

Lemma C.3 *For an arbitrary $b \geq b_0$, define*

$$Z_n(t) = \sup_{\Delta \in \Gamma(t)} \big| V_{nb}(\Delta) - E(V_{nb}(\Delta)) \big|, \tag{C.8}$$

Then there exist a positive constant c^* which does not depend on (n, p, t) such that on the event A_{n1},

$$P\big(Z_n(t) \geq c^* t \delta \sqrt{\log p/n}\big) \quad \leq \quad \exp\big(-\frac{c^{*2} t^2}{16 b^2} \log p\big).$$

Proof. First, we note that by (C.2) and (C.3), $\forall\ \Delta \in \Gamma(t)$,

$$0 \leq |\epsilon_i| \varphi_{\epsilon_i}(X_i^T \Delta) \nu_{b\,||\Delta||_2}(X_i^T \Delta) \leq |\epsilon_i|\, \mathrm{I}\big(|X_i^T \Delta| > |\epsilon_i|\big)\, \mathrm{I}\big(|X_i^T \Delta| \leq b\delta\big) \ \leq\ b\delta.$$

If we change one observation of the sample, the value of $Z_n(t)$ changes at most $2b\delta/n$. By the bounded difference inequality, $\forall\ s > 0$,

$$P\big(Z_n(t) - \mathrm{E}(Z_n(t)) \geq s\big) \leq \exp\big(-\frac{ns^2}{4b^2\delta^2}\big). \tag{C.9}$$

Next, we evaluate $\mathrm{E}(Z_n(t))$. Let $\{\pi_1, ..., \pi_n\}$ be a Rademacher sequence of Bernoulli random variables independent of (X_i, ϵ_i). Let $g_\epsilon(\xi) = |\epsilon| \varphi_\epsilon(\xi) \nu_{b\delta}(\xi)$ be the function defined in Lemma C.1. We have

$$
\begin{aligned}
& \mathrm{E}(Z_n(t)) \\
\leq\ & 2\,\mathrm{E}_{(\pi,X,\epsilon)}\big(\sup_{\Delta\in\Gamma(t)} n^{-1}\sum_{i=1}^{n}\pi_i g_{\epsilon_i}(X_i^T\Delta)\big) \qquad \text{(symmetrization)} \\
\leq\ & 12\,\mathrm{E}_{(X,\epsilon)}\mathrm{E}_{\pi|(X,\epsilon)}\big(\sup_{\Delta\in\Gamma(t)} n^{-1}\sum_{i=1}^{n}\pi_i X_i^T\Delta\big) \qquad \text{(contraction)} \\
\leq\ & 12\sup_{\Delta\in\Gamma(t)}||\Delta||_1\mathrm{E}_X\mathrm{E}_{\pi|X}\big(n^{-1}||\sum_{i=1}^{n}\pi_i X_i||_\infty\big) \\
\leq\ & 12t\delta\mathrm{E}_X\mathrm{E}_{\pi|X}\big(n^{-1}||\sum_{i=1}^{n}\pi_i X_i||_\infty\big).
\end{aligned} \tag{C.10}
$$

where the first inequality follows from the symmetrization theorem (van der Vaart and Wellner (1996)); the second inequality applies lemma C.1 and the contraction theorem (Ledoux and Talagrand (2013)). Next, we will evaluate $\mathrm{E}_{\pi|X}\big(n^{-1}||\sum_{i=1}^{n}\pi_i X_i||_\infty\big)$. We observe that for $1\leq j\leq p$, conditional on X_i, $n^{-1}\sum_{i=1}^{n}\pi_i x_{ij}$ is mean-zero sub-Gaussian with parameter bounded by $n^{-1}\big(\sum_{i=1}^{n}x_{ij}^2\big)^{1/2}$. By the property of sub-Gaussian random variables (Lemma 17.5, van de Geer (2016)), we have

$$
\begin{aligned}
\mathrm{E}_{\pi|X}\Big(\max_{1\leq j\leq p} n^{-1}\big|\sum_{i=1}^{n}\pi_i x_{ij}\big|\Big) &\leq \sqrt{2\log(2p)/n}\max_{1\leq j\leq p}\big(n^{-1}\sum_{i=1}^{n}x_{ij}^2\big)^{1/2} \\
&\leq \sqrt{2m_x\log(2p)/n}
\end{aligned}
$$

on the event A_{n1}. Hence on A_{n1},

$$
\mathrm{E}(Z_n(t)) \leq \frac{1}{2}c^* t\delta\sqrt{\log p/n}, \tag{C.11}
$$

where $c^*=48\sqrt{m_x}$. Taking $s=\frac{1}{2}c^*t\delta\sqrt{\log p/n}$ in (C.9), we have that on A_{n1},

$$
P\big(Z_n(t)\geq c^*t\delta\sqrt{\log p/n}\big) \;\leq\; \exp\big(-\frac{c^{*2}t^2}{16b^2}\log p\big).
$$

□

Proof of Theorem 1. The proof consists of two steps. At the first step, we will establish that the lower bound stated in the theorem holds with high probability for $U_n(\Delta)$ defined in (C.1), which is $S_n(\beta^* + \Delta) - S_n(\beta^*)$ corresponding to a specific choice of subgradient in $\partial Q_n(\beta)$. At the second step, we will show that the lower bound holds for an arbitrary choice of subgradient in $\partial Q_n(\beta)$.

Step 1. Note that $U_n(\Delta) \geq V_{nb}(\Delta)$, $\forall\ b > 0$. We will first prove that, for an arbitrary $b \geq b_0$, there exist some positive constants a_1 and a_2 such that with probability at least $1 - \delta_n - a_1 \exp(-a_2 \log p)$,

$$V_{nb}(\Delta) \geq a^*||\Delta||_2^2 - c^*||\Delta||_1\sqrt{\frac{\log p}{n}} \tag{C.12}$$

uniformly over all $\Delta \in S_2(\delta) \cap \mathbb{C}$, where c^* is the positive constant in Lemma C.3, a_1 and a_2 (specified later in this proof) only depend on the probability distributions of X_i and ϵ_i. Recalling that $S_2(\delta) = \{\Delta : ||\Delta||_2 = \delta\}$, $0 < \delta \leq 1$. Since $0 < \delta \leq 1$, we have $\frac{||\Delta||_1}{||\Delta||_2} \geq 1$. Noting that $V_n(\Delta)$ is always nonnegative, we only need to verify (C.12) for Δ satisfying $a^*||\Delta||_2^2 - c^*||\Delta||_1\sqrt{\frac{\log p}{n}} \geq 0$. It is sufficient to consider $\Delta \in \widetilde{S}_2(\delta)$, where

$$\widetilde{S}_2(\delta) = \Big\{\Delta : \Delta \in S_2(\delta) \cap \mathbb{C},\ 1 \leq \frac{||\Delta||_1}{||\Delta||_2} \leq \frac{a^*}{c^*}\delta\sqrt{\frac{n}{\log p}}\Big\}.$$

Lemma C.2 ensures that $\mathrm{E}(V_{nb}(\Delta)) \geq a^*||\Delta||_2^2$ for any $\Delta \in \widetilde{S}_2(\delta)$. We verify (C.12) by proving

$$\big|V_{nb}(\Delta) - \mathrm{E}(V_{nb}(\Delta))\big| \leq c^*\delta\frac{||\Delta||_1}{||\Delta||_2}\sqrt{\log p/n} \tag{C.13}$$

holds uniformly for $\Delta \in \widetilde{S}_2(\delta)$ with probability at least $1 - a_1 \exp(-a_2 \log p)$ for some positive constants a_1 and a_2. We will prove (C.13) by employing a peeling technique (see, e.g., van der Vaart and Wellner (1996), van de Geer (2000) and the references therein). Write

$$h(\Delta) = \frac{||\Delta||_1}{||\Delta||_2} \quad \text{and} \quad g(h(\Delta)) = c^*\delta h(\Delta)\sqrt{\log p/n}.$$

Define

$$B_m = \{\Delta : 2^{m-1}\mu \le g(h(\Delta)) \le 2^m\mu\} \cap \widetilde{S}_2(\delta), \quad m = 1, ..., M,$$

where $\mu = c^*\delta\sqrt{\log p/n}$, and M is taken as the smallest positive integer such that $2^M \ge \frac{a^*\delta}{c^*}\sqrt{n/\log p}$. Note that if $\Delta \in B_m$, then $2^{m-1} \le h(\Delta) \le 2^m$, $m = 1, ..., M$.

Let $E = \{$the event in $(C.13)$ holds for all $\Delta \in \widetilde{S}_2(\delta)\}$. Then on the event A_{n1}, we have

$$\begin{aligned}
P(E^c) &\le \sum_{m=1}^{M} P\Big(\exists\ \Delta \in \widetilde{S}_2(\delta) \cap B_m \text{ such that } |V_{nb}(\Delta) - \mathrm{E}(V_{nb}(\Delta))| > g(h(\Delta))\Big) \\
&\le \sum_{m=1}^{M} P\Big(\sup_{\Delta \in S_2(\delta)\cap\mathbb{C},\ 2^{m-1} \le h(\Delta) \le 2^m} |V_{nb}(\Delta) - \mathrm{E}(V_{nb}(\Delta))| > c^*\delta 2^{m-1}\sqrt{\log p/n}\Big) \\
&\le \sum_{m=1}^{M} \exp\Big(- a_2 2^{2m-2}\log p\Big) \qquad (\text{by Lemma C.3, taking } a_2 = c^{*2}/(16b^2)) \\
&\le \sum_{m=1}^{M} \exp\Big(- m a_2 \log p\Big) \qquad (\text{noting } m \le 2^{2m-2},\ m = 1, 2, \ldots) \\
&\le \exp(-a_2\log p)\sum_{m=0}^{\infty}\big(\exp(-a_2\log p)\big)^m \\
&= \frac{\exp(-a_2\log p)}{1 - \exp(-a_2\log p)} \qquad (\text{sum of a geometric series}) \\
&\le a_1\exp(-a_2\log p),
\end{aligned}$$

where $a_1 = 1 + \exp(-a_2\log 2)$. Note that $P(A_{n1}) \ge 1 - \delta_n$ for $\delta_n \to 0$. Hence, with probability at least $1 - \delta_n - a_1\exp(-a_2\log p)$, $U_n(\Delta) \ge a^*||\Delta||_2^2 - c^*||\Delta||_1\sqrt{\frac{\log p}{n}}$ uniformly over all $\Delta \in S_2(\delta) \cap \mathbb{C}$.

Step 2. We now consider an arbitrary subgradient $S_n(\beta) = (S_{n1}(\beta), ..., S_{np}(\beta))^T$ in $\partial Q_n(\beta)$. According to (6), $S_j(\beta)$, $j = 1, \ldots, p$, has the form

$$\begin{aligned}
S_{nj}(\beta) &= -\tau n^{-1}\sum_{i=1}^{n} x_{ij} + n^{-1}\sum_{i=1}^{n} x_{ij}\mathrm{I}(Y_i - X_i^T\beta \le 0) \\
&\quad - n^{-1}\sum_{i=1}^{n}[v_i + (1-\tau)]x_{ij}\mathrm{I}(Y_i - X_i^T\beta = 0),
\end{aligned}$$

where $v_i = 0$ if $Y_i - X_i^T\beta = 0$, and $v_i \in [\tau - 1, \tau]$ otherwise, $i = 1, \ldots, n$. Hence,

$$\begin{aligned}
& \langle S_n(\beta^* + \Delta) - S_n(\beta^*), \Delta\rangle \\
\geq\ & n^{-1}\sum_{i=1}^n X_i^T\Delta[\ \mathrm{I}(\epsilon_i \leq X_i^T\Delta) - \ \mathrm{I}(\epsilon_i \leq 0)] - n^{-1}\sum_{i=1}^n |X_i^T\Delta|\ \mathrm{I}(\epsilon_i = X_i^T\Delta) \\
& -n^{-1}\sum_{i=1}^n |X_i^T\Delta|\ \mathrm{I}(\epsilon_i = 0) \\
=\ & U_n(\Delta) - n^{-1}\sum_{i=1}^n |X_i^T\Delta|\ \mathrm{I}(\epsilon_i = X_i^T\Delta) - n^{-1}\sum_{i=1}^n |X_i^T\Delta|\ \mathrm{I}(\epsilon_i = 0).
\end{aligned}$$

Consider any Δ such that $||\Delta||_2 \leq 1$. Consider the event $G_n = \{U_n(\Delta) < k_0||\Delta||_2^2 - c^*||\Delta||_1\sqrt{\log p/n}\}$. We have

$$\begin{aligned}
& P\Big(\ \langle S_n(\beta) - S_n(\beta^*), \Delta\rangle < k_0||\Delta||_2^2 - c^*||\Delta||_1\sqrt{\log p/n}\Big) \\
\leq\ & P\Big(U_n(\Delta) - n^{-1}\sum_{i=1}^n |X_i^T\Delta|\big(\ \mathrm{I}(\epsilon_i = X_i^T\Delta) + \ \mathrm{I}(\epsilon_i = 0)\big) \\
& < k_0||\Delta||_2^2 - c^*||\Delta||_1\sqrt{\log p/n}, G_n^c\Big) + P\big(G_n\big) \\
\leq\ & P\big(n^{-1}\sum_{i=1}^n |X_i^T\Delta|\big(\ \mathrm{I}(\epsilon_i = X_i^T\Delta) + \ \mathrm{I}(\epsilon_i = 0)\big) > 0\big) + P(G_n) \\
\leq\ & 0 + \delta_n + a_1\exp(-a_2\log p) = \delta_n + a_1\exp(-a_2\log p),
\end{aligned}$$

where the last inequality follows because ϵ_i is assumed to have a continuous density (e.g., Lemma A.1 of Ruppert and Carroll (1980)). This proves the theorem. $\square$

We first present a lemma that shows the uniform lower bound in Theorem 1 on $\{||\Delta||_2 \leq 1\} \cap \mathbb{C}$ can be extended to $\{||\Delta||_2 > 1\} \cap \mathbb{C}$.

Lemma C.4 *Suppose conditions (C1)-(C3) are satisfied. There exist some positive constants a^*, c^*, a_1 and a_2, such that for any subgradient $S_n \in \partial Q_n(\beta)$, with probability at least $1 - \delta_n - a_1\exp(-a_2\log p)$,*

$$\langle S_n(\beta) - S_n(\beta^*), \Delta\rangle \geq a^*||\Delta||_2 - c^*\sqrt{\frac{\log p}{n}}||\Delta||_1, \tag{C.14}$$

uniformly on $\{||\Delta||_2 > 1\} \cap \mathbb{C}$, *where the positive constants* a^*, c^*, a_1 *and* a_2 *are those defined in Theorem 1.*

Proof. Observing that $\forall\ 0 < s < 1$, $(X_i^T\Delta)\big[\mathrm{I}(\epsilon_i \leq X_i^T\Delta) - \mathrm{I}(\epsilon_i \leq 0)\big] \geq (sX_i^T\Delta)\big[\mathrm{I}(\epsilon_i \leq X_i^T s\Delta) - \mathrm{I}(\epsilon_i \leq 0)\big] \geq 0$, we have

$$\begin{aligned} U_n(\Delta) &= n^{-1}\sum_{i=1}^n X_i^T\Delta[\mathrm{I}(\epsilon_i \leq X_i^T\Delta) - \mathrm{I}(\epsilon_i \leq 0)] \\ &\geq \frac{1}{s}n^{-1}\sum_{i=1}^n X_i^T(s\Delta)[\mathrm{I}(\epsilon_i \leq X_i^T s\Delta) - \mathrm{I}(\epsilon_i \leq 0)] \\ &= \frac{1}{s}U_n(s\Delta). \end{aligned} \tag{C.15}$$

For $||\Delta||_2 > 1$, we take $s = \frac{1}{||\Delta||_2} \in (0, 1)$ in (C.15) to rescale Δ to $\frac{\Delta}{||\Delta||_2}$ and obtain

$$\begin{aligned} U_n(\Delta) &\geq ||\Delta||_2 U_n\Big(\frac{\Delta}{||\Delta||_2}\Big) \\ &\geq ||\Delta||_2\Big(a^* - c^*\frac{||\Delta||_1}{||\Delta||_2}\sqrt{\frac{\log p}{n}}\Big) \\ &\geq a^*||\Delta||_2 - c^*||\Delta||_1\sqrt{\frac{\log p}{n}}, \end{aligned} \tag{C.16}$$

with probability at least $1 - \delta_n - a_1\exp(-a_2\log p)$, where the second inequality applies Theorem 1. And arguing the same way as in the proof of Theorem 1, we can show that the above lower bound holds for an arbitrary subgradient $S_n(\beta)$ in $\partial Q_n(\beta)$. $\square$

Proof of Theorem 2 (QR-LASSO). Let $\widehat{\beta}$ be the QR-LASSO estimator. Then there exists some subgradient $S_n(\beta) \in \partial Q_n(\beta)$ such that

$$\big\langle S_n(\widehat{\beta}) + \lambda\mathrm{sgn}(\widehat{\beta}),\ \beta^* - \widehat{\beta}\big\rangle = 0, \tag{C.17}$$

where $\mathrm{sgn}(\widehat{\beta}) = (0, \mathrm{sgn}(\widehat{\beta}_2), \ldots, \mathrm{sgn}(\widehat{\beta}_p))^T$.

Let $\widehat{v} = \widehat{\beta} - \beta^*$. We will first use proof by contradiction to show that

$$P\big(||\widehat{v}||_2 \leq 1\big) \geq 1 - 2\delta_n - 2\exp(-\log p) - a_1\exp(-a_2\log p).$$

Let $\Lambda_n = \{\lambda \geq 2||\widetilde{S}_n||_\infty\}$. The choice of λ and Lemma 3 imply that $P(\Lambda_n) \geq 1 - \delta_n - 2\exp(-\log p)$. Assume $||\widehat{v}||_2 > 1$. Define the event $B_{n1} = \{\langle S_n(\widehat{\beta}) - S_n(\beta^*), \widehat{v}\rangle \geq a^*||\widehat{v}||_2 - c^*||\widehat{v}||_1\sqrt{\log p/n}\}$. If $||\widehat{v}||_2 > 1$, Lemma C.4 implies that $P(B_{n1}) \geq 1 - \delta_n - a_1\exp(-a_2\log p)$. It follows that $P(\Lambda_n \cap B_{n1}) \geq 1 - 2\delta_n - 2\exp(-\log p) - a_1\exp(-a_2\log p)$. It is sufficient to show that a contradiction occurs on the event $\Lambda_n \cap B_{n1}$.

To see the contradiction, we first observe that on B_{n1}, (C.17) implies that

$$\langle -\lambda\text{sgn}(\widehat{\beta}) - S_n(\beta^*),\ \widehat{v}\rangle \geq a^*||\widehat{v}||_2 - c^*||\widehat{v}||_1\sqrt{\log p/n}. \tag{C.18}$$

On the other hand, by Hölder's inequality,

$$\langle -\lambda\text{sgn}(\widehat{\beta}) - S_n(\beta^*),\ \widehat{v}\rangle \ \leq \ \{\lambda + ||S_n(\beta^*)||_\infty\}||\widehat{v}||_1.$$

As the argument in step 2 of the proof of Theorem 1 implies, for an arbitrary subgradient $S_n(\cdot)$, on the event Λ_n, $||S_n(\beta^*)||_\infty \leq 2\lambda$ with probability one. Hence, on Λ_n,

$$\langle -\lambda\text{sgn}(\widehat{\beta}) - S_n(\beta^*),\ \widehat{v}\rangle \ \leq \ 1.5k_0\sqrt{\log p/n}||\widehat{v}||_1. \tag{C.19}$$

(C.18) and (C.19) together imply that

$$a^*||\widehat{v}||_2 \leq (2k_0 + c^*)\sqrt{\log p/n}||\widehat{v}||_1. \tag{C.20}$$

(i) (Hard-sparsity case) The proof of Lemma 2 ensures that on the event Λ_n we have $\widehat{v} \in \Gamma_H$ and hence $||\widehat{v}||_1 \leq 4\sqrt{s}||\widehat{v}||_2$. Then by (C.20), $a^* \leq 4(2k_0 + c^*)\sqrt{s\log p/n}$. This contradicts with the assumption $n > c's\log p$, where $c' = 16(2k_0 + c^*)^2/(a^*)^2$.

(ii) (Soft-sparsity case) The proof of Lemma 2 ensures that on the event Λ_n we have $\widehat{v} \in \Gamma_W = \{v \in \mathbb{R}^p : ||v_{S_a^c}||_1 \leq 3||v_{S_a}||_1 + 4||\beta^*_{S_a^c}||_1\}$, where $S_a = S_{-a} \cup \{1\}$ with the index set $S_{-a} = \{j : |\beta^*_j| > a, 2 \leq j \leq p\}$ and S_a^c denotes the complement of S_a in $\{1, 2\ldots, p\}$. Under the soft sparsity assumption, we have $a|S_{-a}| \leq \sum_{i=2}^p |\beta^*_i| \leq R$, where $|S_{-a}|$ denotes the cardinality of the set S_{-a}. Hence $|S_{-a}| \le a^{-1}R$. Since $\widehat{v} \in \Gamma_W$, we have

$$||\widehat{v}||_1 \;\; \le \;\; 4\big(||\widehat{v}_{S_a}||_1 + ||\beta^*_{S_a^c}||_1\big) \le 4\big(\sqrt{a^{-1}R + 1}||\widehat{v}||_2 + R\big), \tag{C.21}$$

which holds for any $a > 0$. Taking $a = R/3$, we obtain

$$||\widehat{v}||_1 \le 4\big[2||\widehat{v}||_2 + R\big]. \tag{C.22}$$

Hence (C.20) contradicts with the assumption $c^{''}\sqrt{\log p/n}\max\{2, R\} < 1/2$, where $c^{''} = 4(2k_0 + c^*)/a^*$.

Define the event $D_{n1} = \big\{||\widehat{v}||_2 \le 1\big\}$. In the above, we have verified that $P(D_{n1}) \ge 1 - 2\delta_n - 2\exp(-\log p) - a_1\exp(-a_2\log p)$. Define $B_{n2} = \big\{\big\langle S_n(\widehat{\beta}) - S_n(\beta^*),\ \widehat{\beta} - \beta^*\big\rangle \ge a^*||\widehat{v}||_2^2 - c^*||\widehat{v}||_1\sqrt{\log p/n}\big\}$. By Theorem 1, on D_{n1}, $P(B_{n2}) \ge 1 - \delta_n - a_1\exp(-a_2\log p)$. From now on, we consider the event $\Lambda_n \cap D_{n1} \cap B_{n1}$. We have $P(\Lambda_n \cap D_{n1} \cap B_{n1}) \ge 1 - 4\delta_n - 4\exp(-\log p) - 2a_1\exp(-a_2\log p)$. By the convexity of $||\beta||_1$, we have

$$||\beta^*||_1 - ||\widehat{\beta}||_1 \;\; \ge \;\; \big\langle \mathrm{sgn}(\widehat{\beta}),\ \beta^* - \widehat{\beta}\big\rangle. \tag{C.23}$$

Combining (C.23) with (C.17), we have

$$\big\langle S_n(\widehat{\beta}),\ \beta^* - \widehat{\beta}\big\rangle = -\big\langle \lambda\mathrm{sgn}(\widehat{\beta}),\ \beta^* - \widehat{\beta}\big\rangle \ge \lambda\big(||\widehat{\beta}||_1 - ||\beta^*||_1\big). \tag{C.24}$$

On B_{n2}, (C.24) implies that

$$\big\langle -S_n(\beta^*),\ \widehat{\beta} - \beta^*\big\rangle \ge a^*||\widehat{v}||_2^2 + \lambda\big(||\widehat{\beta}||_1 - ||\beta^*||_1\big) - c^*||\widehat{v}||_1\sqrt{\log p/n}.$$

Applying Hölder's inequality, we have

$$a^*||\widehat{v}||_2^2 + \lambda\big(||\widehat{\beta}||_1 - ||\beta^*||_1\big) - c^*||\widehat{v}||_1\sqrt{\log p/n} \le ||S_n(\beta^*)||_\infty||\widehat{v}||_1.$$

Rearranging terms and applying the triangle inequality, we obtain

$$\begin{aligned} a^*||\widehat{v}||_2^2 &\le \big(\lambda + c^*\sqrt{\log p/n} + ||S_n(\beta^*)||_\infty\big)||\widehat{v}||_1 \\ &\le (2k_0 + c^*)\sqrt{\log p/n}||\widehat{v}||_1. \end{aligned} \tag{C.25}$$

(i) For the hard-sparsity case, on Λ_n, $||\widehat{v}||_1 \le 4\sqrt{s}||\widehat{v}||_2$. Then (C.25) implies

$$||\widehat{v}||_2 \le a_1^*\sqrt{s\log p/n},$$

where $a_1^* = 4(2k_0 + c^*)/a^*$. We also obtain $||\widehat{v}||_1 \le 4a_1^* s\sqrt{\log p/n}$.

(ii) For the soft-sparsity case, on Λ_n, (C.25) and (C.21) imply that

$$||\widehat{v}||_2 \le 2\max\left\{\frac{4(2k_0+c^*)}{a^*}\sqrt{a^{-1}R+1}\sqrt{\log p/n}, 2\sqrt{\frac{2k_0+c^*}{a^*}R^{1/2}(\log p/n)^{1/4}}\right\},$$

for any $a > 0$. Taking $a = \sqrt{\log p/n}$, we have

$$\sqrt{a^{-1}R+1}\sqrt{\log p/n} = \sqrt{R + (\log p/n)^{1/2}}(\log p/n)^{1/4} \le \sqrt{2R}(\log p/n)^{1/4}.$$

We obtain

$$||\widehat{v}||_2 \le a_2^* R^{1/2}(\log p/n)^{1/4},$$

where $a_2^* = 4\max\left\{2\sqrt{2}(2k_0+c^*)/a^*, \sqrt{(2k_0+c^*)/a^*}\right\}$. The same reasoning that leads to (C.22) also implies that

$$\begin{aligned} ||\widehat{v}||_1 &\le 4\big(\sqrt{|S_a|}||\widehat{v}||_2 + ||\beta^*_{S_a^c}||_1\big) \\ &\le 4\big(a_2^*\sqrt{|S_a|}R^{1/2}(\log p/n)^{1/4} + ||\beta^*_{S_a^c}||_1\big) \end{aligned}$$

for any $a > 0$, where $|S_a|$ denotes the cardinality of the set S_a. □

Proof of Theorem 3 (QR-NCP). The proof is based on the same idea as in the proof of Theorem 2 but is more involved. We consider any feasible local solution $\widehat{\beta}$ such that

$||\widehat{\beta}||_1 \leq \kappa$ and (8) is satisfied. Then there exists some subgradient $S_n(\beta) \in \partial Q_n(\beta)$ such that

$$\big\langle S_n(\widehat{\beta}) + \lambda \mathrm{sgn}(\widehat{\beta}) - H'(\widehat{\beta}),\ \beta^* - \widehat{\beta}\big\rangle = 0, \tag{C.26}$$

where $H'(\widehat{\beta}) = (0, h'_\lambda(\widehat{\beta}_2), \ldots, h'_\lambda(\widehat{\beta}_p))^T$ and $\mathrm{sgn}(\widehat{\beta}) = (0, \mathrm{sgn}(\widehat{\beta}_2), \ldots, \mathrm{sgn}(\widehat{\beta}_p))^T$. Denote $\widehat{v} = \widehat{\beta} - \beta^*$.

Let $\widetilde{\Lambda}_n = \big\{\lambda \geq 4||\widetilde{S}_n||_\infty\big\}$, where λ satisfies the condition of Theorem 3. Lemma 3 imply that $P(\widetilde{\Lambda}_n) \geq 1 - \delta_n - 2\exp(-\log p)$. Let $B_{n1} = \big\{\big\langle S_n(\widehat{\beta}) - S_n(\beta^*), \widehat{v}\big\rangle \geq a^*||\widehat{v}||_2 - c^*||\widehat{v}||_1\sqrt{\log p/n}\big\}$. If $||\widehat{v}||_2 > 1$, Lemma C.4 implies that $P(B_{n1}) \geq 1 - \delta_n - a_1\exp(-a_2\log p)$. We have $P(\widetilde{\Lambda}_n \cap B_{n1}) \geq 1 - 2\delta_n - 2\exp(-\log p) - a_1\exp(-a_2\log p)$. We will first use proof by contradiction to show that

$$P\big(||\widehat{v}||_2 \leq 1\big) \geq 1 - 2\delta_n - 2\exp(-\log p) - a_1\exp(-a_2\log p).$$

Assume $||\widehat{v}||_2 > 1$. It is sufficient to show that a contradiction occurs on the event $\widetilde{\Lambda}_n \cap B_{n1}$.

On the event B_{n1}, by (C.26), we have

$$\big\langle H'(\widehat{\beta}) - \lambda\mathrm{sgn}(\widehat{\beta}) - S_n(\beta^*),\ \widehat{v}\big\rangle \geq a^*||\widehat{v}||_2 - c^*||\widehat{v}||_1\sqrt{\frac{\log p}{n}} \tag{C.27}$$

On the other hand, by Hölder's inequality,

$$\big\langle H'(\widehat{\beta}) - \lambda\mathrm{sgn}(\widehat{\beta}) - S_n(\beta^*),\ \widehat{v}\big\rangle \ \leq\ \big\{||H'(\widehat{\beta}) - \lambda\mathrm{sgn}(\widehat{\beta})||_\infty + ||S_n(\beta^*)||_\infty\big\}||\widehat{v}||_1$$

Note that $-H(\beta_-) + \lambda||\beta_-||_1 = p_\lambda(\beta_-) = \sum_{j=2}^p p_\lambda(|\beta_j|)$. We have $||H'(\widehat{\beta}) - \lambda\mathrm{sgn}(\widehat{\beta})||_\infty = ||\partial p_\lambda(\widehat{\beta}_-)||_\infty$, which is upper bounded by λ (see, e.g., Lemma 4, Loh and Wainwright (2015)). Furthermore, on $\widetilde{\Lambda}_n$, $||S_n(\beta^*)||_\infty \leq \lambda/4$. Hence,

$$\big\langle H'(\widehat{\beta}) - \lambda\mathrm{sgn}(\widehat{\beta}) - S_n(\beta^*),\ \widehat{v}\big\rangle \ \leq\ \frac{5\lambda}{4}||\widehat{v}||_1. \tag{C.28}$$

By the choice of λ, (C.27) and (C.28) together imply that

$$a^*||\widehat{v}||_2 \leq \big(5\lambda/4 + c^*\sqrt{\log p/n}\big)||\widehat{v}||_1 \leq \frac{3}{2}\lambda||\widehat{v}||_1 \leq \frac{3\kappa k_0}{2}\sqrt{\log p/n}.$$

Hence we have a contradiction under the assumption $\sqrt{\log p/n} < \frac{2a^*}{3\kappa k_0}$.

Define the event $D_{n1} = \big\{||\widehat{v}||_2 \le 1\big\}$. In the above, we have verified that $P(D_{n1}) \ge 1 - 2\delta_n - 2\exp(-\log p) - a_1\exp(-a_2\log p)$. Define $B_{n2} = \big\{\big\langle S_n(\widehat{\beta}) - S_n(\beta^*),\ \widehat{\beta} - \beta^*\big\rangle \ge a^*||\widehat{v}||_2^2 - c^*||\widehat{v}||_1\sqrt{\log p/n}\big\}$. By Theorem 1, on D_{n1}, $P(B_{n2}) \ge 1 - \delta_n - a_1\exp(-a_2\log p)$. From now on, we consider the event $\widetilde{\Lambda}_n \cap D_{n1} \cap B_{n1}$. We have $P(\widetilde{\Lambda}_n \cap D_{n1} \cap B_{n1}) \ge 1 - 4\delta_n - 4\exp(-\log p) - 2a_1\exp(-a_2\log p)$.

By the assumption on the penalty function, $\frac{\gamma_0}{2}||\beta||_2^2 - H(\beta_-) + \lambda||\beta_-||_1$ is convex in β. This convexity property implies that

$$\begin{aligned}&\big[\frac{\gamma_0}{2}||\beta^*||_2^2 - H(\beta^*_-) + \lambda||\beta^*_-||_1\big] - \big[\frac{\gamma_0}{2}||\widehat{\beta}||_2^2 - H(\widehat{\beta}_-) + \lambda||\widehat{\beta}_-||_1\big]\\ \ge\ &\big\langle \gamma_0\widehat{\beta} - H'(\widehat{\beta}) + \lambda\mathrm{sgn}(\widehat{\beta}),\ \beta^* - \widehat{\beta}\big\rangle.\end{aligned}$$

So we have

$$\begin{aligned}&p_\lambda(\beta^*_-) - p_\lambda(\widehat{\beta}_-) + \frac{\gamma_0}{2}\big(||\beta^*||_2^2 - ||\widehat{\beta}||_2^2\big)\\ \ge\ &\big\langle -H'(\widehat{\beta}) + \lambda\mathrm{sgn}(\widehat{\beta}),\ \beta^* - \widehat{\beta}\big\rangle + \gamma_0\big\langle\widehat{\beta}, \beta^*\big\rangle - \gamma_0||\widehat{\beta}||_2.\end{aligned}$$

Or equivalently,

$$p_\lambda(\beta^*_-) - p_\lambda(\widehat{\beta}_-) + \frac{\gamma_0}{2}||\beta^* - \widehat{\beta}||_2^2 \ge \big\langle -H'(\widehat{\beta}) + \lambda\mathrm{sgn}(\widehat{\beta}),\ \beta^* - \widehat{\beta}\big\rangle. \tag{C.29}$$

It follows from (C.26) and (C.29) that

$$\begin{aligned}&\big\langle S_n(\widehat{\beta}),\ \beta^* - \widehat{\beta}\big\rangle\\ =\ &-\big\langle -H'(\widehat{\beta}) + \lambda\mathrm{sgn}(\widehat{\beta}),\ \beta^* - \widehat{\beta}\big\rangle \ge p_\lambda(\widehat{\beta}_-) - p_\lambda(\beta^*_-) - \frac{\gamma_0}{2}||\widehat{v}||_2^2.\end{aligned} \tag{C.30}$$

By Theorem 1, on B_{n2},

$$\big\langle S_n(\widehat{\beta}) - S_n(\beta^*),\ \widehat{\beta} - \beta^*\big\rangle \ge a^*||\widehat{v}||_2^2 - c^*||\widehat{v}||_1\sqrt{\log p/n}. \tag{C.31}$$

(C.30) and (C.31) together imply that

$$\big\langle -S_n(\beta^*),\ \widehat{v}\big\rangle \geq \big(a^* - \gamma_0/2\big)||\widehat{v}||_2^2 + \big(p_\lambda(\widehat{\beta}_-) - p_\lambda(\beta^*_-)\big) - c^*||\widehat{v}||_1\sqrt{\log p/n}.$$

Applying Hölder's inequality, we have

$$\big(a^* - \gamma_0/2\big)||\widehat{v}||_2^2 + \big(p_\lambda(\widehat{\beta}_-) - p_\lambda(\beta^*_-)\big) - c^*||\widehat{v}||_1\sqrt{\log p/n} \leq ||S_n(\beta^*)||_\infty ||\widehat{v}||_1.$$

Hence,

$$\big(a^* - \gamma_0/2\big)||\widehat{v}||_2^2 \leq (p_\lambda(\beta^*_-) - p_\lambda(\widehat{\beta}_-)) + \big(c^*\sqrt{\log p/n} + ||S_n(\beta^*)||_\infty\big)||\widehat{v}||_1.$$

On $\widetilde{\Lambda}_n$, we have $||S_n(\beta^*)||_\infty \leq \frac{\lambda}{4}$, and the choice of λ implies $c^*\sqrt{\log p/n} \leq \frac{\lambda}{4}$. We hence have

$$\big(a^* - \gamma_0/2\big)||\widehat{v}||_2^2 \ \ \leq \ \ \big(p_\lambda(\beta^*_-) - p_\lambda(\widehat{\beta}_-)\big) + \frac{\lambda}{2}||\widehat{v}||_1. \tag{C.32}$$

(i) Hard-sparsity case. Write $\widehat{v} = (\widehat{v}_1, \widehat{v}_-^T)^T$, where $\widehat{v}_1 = \widehat{\beta}_1 - \beta^*_1$ corresponding to the intercept term, and $\widehat{v}_- = \widehat{\beta}_- - \beta^*_-$. Recall $S_- = \{j : \beta^*_j \neq 0,\ 2 \leq j \leq p\}$, $S = S_- \cup \{1\}$, and $||S||_0 = s$ is the model sparsity size under the hard sparsity condition. Let β^*_{S-} denote the $(s-1)$-subvector of β^* consisting of the elements in S_-. Let $\beta^*_{S^c}$, $\widehat{v}_{S_-}$ and $\widehat{v}_{S^c}$ be defined similarly. Note that under the hard-sparsity assumption, $\beta^*_{S^c}$ is a $(p-s)$-dimensional zero vector. By the sub-additivity property of the penalty function $p_\lambda(\cdot)$, we have $p_\lambda(|t_1+t_2|) \geq p_\lambda(|t_1|) - p_\lambda(|t_2|)$ for any $t_1, t_2 \in \mathbb{R}$, due to the observation $|t_1| \leq |t_2| + |t_1 + t_2|$. Applying the sub-additivity property, we have

$$p_\lambda(\beta^*_-) - p_\lambda(\widehat{\beta}_-) \leq p_\lambda(\beta^*_{S_-}) - \big[p_\lambda(\beta^*_{S_-} + \widehat{v}_{S^c}) - p_\lambda(\widehat{v}_{S_-})\big] = p_\lambda(\widehat{v}_{S_-}) - p_\lambda(\widehat{v}_{S^c}). \tag{C.33}$$

By the assumption on the penalty function and Lemma 4 of Loh and Wainwright (2015), we have

$$\lambda||\widehat{v}||_1 \leq p_\lambda(\widehat{v}) + \frac{\gamma_0}{2}||\widehat{v}||_2^2 \leq p_\lambda(\widehat{v}_{S_-}) + p_\lambda(\widehat{v}_{S^c}) + p_\lambda(\widehat{v}_1) + \frac{\gamma_0}{2}||\widehat{v}||_2^2. \tag{C.34}$$

Combining (C.32) with (C.33) and (C.34), we have

$$\big(a^* - 3\gamma_0/4\big)||\widehat{v}||_2^2 \leq \frac{3}{2}p_\lambda(\widehat{v}_{S_-}) - \frac{1}{2}p_\lambda(\widehat{v}_{S^c}) + \frac{1}{2}p_\lambda(\widehat{v}_1) \leq \frac{3}{2}p_\lambda(\widehat{v}_S), \tag{C.35}$$

where $\widehat{v}_S = (\widehat{v}_1, \widehat{v}_{S_-}^T)^T$. As $p_\lambda(t)$ is a λ-Lipschitz continuous function of t (Lemma 4 of Loh and Wainwright (2015)), we have $p_\lambda(\widehat{v}_S) \leq \lambda||\widehat{v}_S|_1$. Hence

$$\big(a^* - 3\gamma_0/4\big)||\widehat{v}||_2^2 \leq \frac{3\lambda}{2}||\widehat{v}_S||_1 \leq \frac{3\lambda}{2}\sqrt{s}||\widehat{v}||_2.$$

We have proved $||\widehat{v}||_2 \leq a_3^*\sqrt{s\log p/n}$, where $a_3^* = \frac{6k_0}{4a^*-3\gamma_0}$. From the above and the argument for Lemma 1, $||v_{A^c}||_1 \leq 3||v_A||_1$, where v_A denotes the subvector of v that contains the s-largest elements of v. Hence, $||\widehat{v}||_1 \leq 4a_3^* s\sqrt{\log p/n}$.

(ii) Soft-sparsity case. Consider $S_a = S_{-a} \cup \{1\}$, where $S_{-a} = \{j : |\beta_j^*| > a, 2 \leq j \leq p\}$ and a is an arbitrary positive constant. Applying the sub-additivity property of the penalty function $p_\lambda(\cdot)$, we have

$$\begin{aligned} & p_\lambda(\beta_-^*) - p_\lambda(\beta_-) \\ \leq\ & \big[p_\lambda(\beta_{S_{-a}}^*) + p_\lambda(\beta_{S_a^c}^*)\big] - \big[p_\lambda(\beta_{S_{-a}}^* + \widehat{v}_{S_a^c}) - p_\lambda(\beta_{S_a^c}^* + \widehat{v}_{S_{-a}})\big] \\ =\ & p_\lambda(\widehat{v}_{S_{-a}}) - p_\lambda(\widehat{v}_{S_a^c}) + p_\lambda(\beta_{S_a^c}^*). \end{aligned} \tag{C.36}$$

And similarly as (C.34), we have

$$\lambda||\widehat{v}||_1 \leq p_\lambda(\widehat{v}) + \frac{\gamma_0}{2}||\widehat{v}||_2^2 \leq p_\lambda(\widehat{v}_{S_{-a}}) + p_\lambda(\widehat{v}_{S_a^c}) + p_\lambda(\widehat{v}_1) + \frac{\gamma_0}{2}||\widehat{v}||_2^2. \tag{C.37}$$

Combining (C.32) with (C.36) and (C.37) and noting $p_\lambda(\beta^*_{S_a^c}) \le \lambda||\beta^*_{S_a^c}||_1 \le \lambda R$, we have

$$\begin{aligned}\big(a^* - 3\gamma_0/4\big)||\widehat{v}||_2^2 &\le \frac{3}{2}p_\lambda(\widehat{v}_{S_{-a}}) - \frac{1}{2}p_\lambda(\widehat{v}_{S_a^c}) + \frac{1}{2}p_\lambda(\widehat{v}_1) + p_\lambda(\beta^*_{S_a^c}) \\ &\le \frac{3}{2}p_\lambda(\widehat{v}_{S_a}) - \frac{1}{2}p_\lambda(\widehat{v}_{S_a^c}) + \lambda||\beta^*_{S_a^c}||_1. \end{aligned} \tag{C.38}$$

As $p_\lambda(\widehat{v}_{S_a^c}) \ge \lambda||\widehat{v}_{S_a^c}||_1 - \frac{\gamma_0}{2}||||\widehat{v}_{S_a^c}||_2^2$, we have

$$\big(a^* - \gamma_0\big)||\widehat{v}||_2^2 \le \frac{3}{2}\lambda||\widehat{v}_{S_a}||_1 - \frac{1}{2}\lambda||\widehat{v}_{S_a^c}||_1 + \lambda||\beta^*_{S_a^c}||_1. \tag{C.39}$$

Similarly as the derivation for (C.21), $||\widehat{v}_{S_a}||_1 \le \sqrt{a^{-1}R+1}||\widehat{v}||_2$. We have

$$\big(a^* - \gamma_0\big)||\widehat{v}||_2^2 \le \frac{3}{2}\sqrt{a^{-1}R+1}\lambda||\widehat{v}||_2 + \lambda R.$$

Hence,

$$||\widehat{v}||_2 \le 2\max\Big\{\frac{3k_0}{2(a^*-\gamma_0)}\sqrt{a^{-1}R+1}\sqrt{\log p/n}, \sqrt{k_0/(a^*-\gamma_0)}R^{1/2}(\log p/n)^{1/4}\Big\}.$$

As in the proof of Theorem 2, taking $a = \sqrt{\log p/n}$ leads to $\sqrt{a^{-1}R+1}\sqrt{\log p/n} \le \sqrt{2R}(\log p/n)^{1/4}$. We then obtain

$$||\widehat{v}||_2 \le a_4^* R^{1/2}(\log p/n)^{1/4},$$

where $a_4^* = 2\max\Big\{\frac{3\sqrt{2}k_0}{2(a^*-\gamma_0)}, \sqrt{\frac{k_0}{a^*-\gamma_0}}\Big\}$. On the event of $\widetilde{\Lambda}_n$, the argument for Lemma 1 ensures that $||\widehat{v}_{S_a^c}||_1 \le 3||\widehat{v}_{S_a}||_1 + 2||\beta^*_{S_a^c}||_1$. Hence,

$$\begin{aligned}||\widehat{v}||_1 &\le 2\big(2\sqrt{|S_a|}||\widehat{v}||_2 + ||\beta^*_{S_a^c}||_1\big) \\ &\le 2\big(2a_4^*\sqrt{|S_a|}R^{1/2}(\log p/n)^{1/4} + ||\beta^*_{S_a^c}||_1\big)\end{aligned}$$

for any $a > 0$, where $|S_a|$ denotes the cardinality of the set S_a. $\square$

Proof of Theorem 4. Inequalities (17) and (18) imply that

$$\begin{aligned}|Q_n(\widehat{\beta}) - Q_n(\beta^*)| &\leq \max\left\{|S_n(\beta^*)^T(\widehat{\beta} - \beta^*)|, |\overline{S}_n(\widehat{\beta})^T(\widehat{\beta} - \beta^*)|\right\} \\ &\leq \max\left\{||S_n(\beta^*)||_\infty, ||\overline{S}_n(\widehat{\beta})||_\infty\right\}||\widehat{\beta} - \beta^*||_1.\end{aligned}$$

The above inequality holds for any subgradients $S_n(\beta), \overline{S}_n(\beta) \in \partial Q_n(\beta)$. We take $S_n(\beta^*) = n^{-1}\sum_{i=1}^n X_i\xi_i$, where $\xi_i = \mathrm{I}(\epsilon_i < 0) - \tau$. By Lemma 3, $P\big(\lambda \geq 4||S_n(\beta^*)||_\infty\big) \geq 1 - \delta_n - 2\exp(-\log p)$. We take $\overline{S}_n(\widehat{\beta})$ to be the subgradient satisfying (8), whose existence is guaranteed by the KKT condition for convex difference program. Then

$$\overline{S}_n(\widehat{\beta}) + \lambda \mathrm{sgn}(\widehat{\beta}) - H'(\widehat{\beta}_-) = 0, \tag{C.40}$$

where $H'(\cdot)$ is defined in Section 2.3 of the main paper. Note that for QR-LASSO, $H'(\cdot) = 0$ and thus $||\overline{S}_n(\widehat{\beta})||_\infty = \lambda$. While for QR-NCP, Note that $-H(\beta) + \lambda||\beta||_1 = p_\lambda(\beta) = \sum_{j=2}^p p_\lambda(|\beta_j|)$. We have $||H'(\widehat{\beta}) - \lambda\mathrm{sgn}(\widehat{\beta})||_\infty = ||\partial p_\lambda(\widehat{\beta}_-)||_\infty$, which is upper bounded by λ (e.g., Lemma 4, Loh and Wainwright (2015)). Hence, $||\overline{S}_n(\widehat{\beta})||_\infty \leq \lambda$. Summarizing the above, we have with probability at least $1 - \delta_n - 2\exp(-\log p)$, $|Q_n(\widehat{\beta}) - Q_n(\beta^*)| \leq 4\lambda||\widehat{\beta} - \beta^*||_1$. $\square$

Appendix D Additional technical results

Lemma D.1 *Suppose Assumption 2 is satisfied. For a positive parameter $k \geq 1$, let $\mathbb{C}(k) = \{\theta \in \mathbb{R}^p : ||\theta||_2 \leq 1, ||\theta||_0 \leq k\}$. Then there exist some positive constants α_1 and α_2 such that $\forall\ t > 0$,*

$$\begin{aligned}&P\Big\{\sup_{\theta\in\mathbb{C}(k)} \big|n^{-1}||\mathbb{X}\theta||_2^2 - E(n^{-1}||\mathbb{X}\theta||_2^2)\big| \geq t\Big\} \\ \leq\ &\alpha_2 \exp\big(-\alpha_1 n \min(t^2/\sigma_x^4, t/\sigma_x^2) + 2k\log p\big).\end{aligned} \tag{D.1}$$

Proof. This is a minor extension of Lemma 15 of Loh and Wainwright (2012) to allow X to include an intercept term. We provide below an outline of the derivation. First we show that the exponential inequality in Lemma 14 of Loh and Wainwright (2012) can be extended to allow X to include an intercept term. Specifically, write $\mathbb{X} = (1_n, \widetilde{\mathbb{X}})$, where 1_n denotes an $n \times 1$ column vector of ones, $\widetilde{\mathbb{X}}$ is an $n \times (p-1)$ matrix of covariates where each row is a sub-Gaussain vector and the rows are independent. For an arbitrary $p \times 1$ vector θ, write $\theta = (\theta_1, \widetilde{\theta}')'$. Then $n^{-1}||\mathbb{X}\theta||_2^2 = \theta_1^2 + n^{-1}||\widetilde{\mathbb{X}}\widetilde{\theta}||_2^2 + 2\theta_1 n^{-1} 1_n' \widetilde{\mathbb{X}}\widetilde{\theta}$. For any $\theta \in \mathbb{R}^p$ and any $t > 0$,

$$\begin{aligned} & P\Big[\big|||\mathbb{X}\theta||_2^2 - E(||\mathbb{X}\theta||_2^2)\big| \geq nt\Big] \\ \leq \; & P\Big[\big|||\widetilde{\mathbb{X}}\widetilde{\theta}||_2^2 - E(||\widetilde{\mathbb{X}}\widetilde{\theta}||_2^2)\big| \geq nt/2\Big] + P\Big[\big|2\theta_1 n^{-1} 1_n' \widetilde{\mathbb{X}}\widetilde{\theta}\big| \geq nt/2\Big]. \end{aligned}$$

Consider any fixed θ such that $||\theta||_2 \leq 1$. By Lemma 14 of Loh and Wainwright (2012), the first term at the right side of the inequality is upper bounded by $2\exp\big(-c_1 n \min(t^2/\sigma_x^4, t/\sigma_x^2)\big)$ for some positive constant c_1. Observing that $n^{-1} 1_n' \widetilde{\mathbb{X}}\widetilde{\theta}$ is an average of i.i.d. sub-Gaussian random variables, the second term at the right side of the inequality is upper bounded by $\exp\big(-c_2 n t^2/\sigma_x^2\big)$ for some positive constant c_2. Therefore,

$$P\Big[\big|||\mathbb{X}\theta||_2^2 - E(||\mathbb{X}\theta||_2^2)\big| \geq nt\Big] \leq c_3 \exp\big(-c_4 n \min(t^2/\sigma_x^4, t/\sigma_x^2)\big), \tag{D.2}$$

for some positive constants c_3 and c_4. Using this exponential inequality and applying the same technique in the proof of Lemma 15 of Loh and Wainwright (2012) establishes the result. □

Lemma D.2 *Assume the conditions of Lemma B.1 are satisfied. We have*

$$P\big(||\widehat{\Sigma}||_\infty \leq 12\zeta_0^2\big) \geq 1 - 2\exp(-\log p).$$

Proof. First, by Cauchy-Schwarz inequality, $|\widehat{\Sigma}_{jk}| = |n^{-1}\sum_{i=1}^n x_{ij}x_{ik}| \leq (\widehat{\Sigma}_{jj})^{1/2}(\widehat{\Sigma}_{jk})^{1/2}$, $\forall\ 1 \leq j,k \leq p$. Hence, $||\widehat{\Sigma}||_\infty \leq \max_{1\leq j\leq p} n^{-1}\sum_{i=1}^n x_{ij}^2$. As in this example, x_{ij} is sub-Gaussian with variance proxy bounded by ζ_0^2, we have $\Sigma_{jj} = \mathrm{E}(x_{ij}^2) \leq 4\zeta_0^2$, $j = 1, \ldots, p$. A mean zero random variable x is sub-exponential with parameter (ζ_*^2, b), denoted by $\mathrm{SE}(\zeta_*^2, b)$, if $\mathrm{E}\{e^{tx}\} \leq \exp(\zeta_*^2 t^2/2)$ for ant $|t| \leq \frac{1}{b}$. The sub-Gaussian property of x_{ij} implies that $x_{ij}^2 - \mathrm{E}(x_{ij}^2) \sim \mathrm{SE}(256\zeta_0^4, 16\zeta_0^2)$, $j = 1, \ldots, p$.

Applying the Bernstein's inequality for subexponential randomm variables, $\forall\, t > 0$,

$$P\Big(n^{-1}\big|\sum_{i=1}^{n}(x_{ij}^2 - \mathrm{E}(x_{ij}^2))\big| > t\Big) \leq 2\exp\Big\{-\frac{n}{2}\min\big(\frac{t^2}{256\zeta_0^4}, \frac{t}{16\zeta_0^2}\big)\Big\}. \tag{D.3}$$

Taking $t = 32\zeta_0^2\sqrt{\log p/n}$ and noting that we assume $\log p \leq n/4$, by the union bound, we have

$$P\big(\max_{1\leq j\leq p}|\widehat{\Sigma}_{jj} - \Sigma_{jj}| > 16\zeta_0^2\sqrt{\log p/n}\big) \leq 2\exp(-\log p).$$

This implies with probability at least $1 - 2\exp(-\log p)$, we have $P(||\widehat{\Sigma}||_\infty \leq 12\zeta_0^2)$ $\square$

Lemma D.3 *Assume the conditions of Lemma B.1 are satisfied. If* $\sqrt{\log p/n} \leq \zeta_1$*, where* $\zeta_1 = \zeta_*/(32e\zeta_0)$*, then*

$$P\Big(\min_{1\leq j\leq p}\sum_{i=1}^{n}|x_{ij}| \geq \zeta_*/2\Big) \geq 1 - 2\exp(-\log p).$$

Proof. We will first verify that if a random variable x is sub-Gaussian with variance proxy ζ_0^2, then $|x|$ is sub-exponential. The sub-Gaussian property of x implies that $\mathrm{E}(|x|^k) \leq (2\zeta_0^2)^{k/2}k\Gamma(k/2)$, for any positive integer $k \geq 1$. We consider below the moment generating function of $|x| - \mathrm{E}(|x|)$. For any $t \in \mathbb{R}$,

$$\begin{aligned}
& \mathrm{E}\Big\{ \exp\big[t(|x| - \mathrm{E}(|x|))\big]\Big\} \\
= \quad & 1 + \sum_{k=2}^{\infty} \frac{t^k \mathrm{E}\big[(|x| - \mathrm{E}(|x|))^k\big]}{k!} \\
\leq \quad & 1 + \sum_{k=2}^{\infty} \frac{t^k 2^{k-1} \mathrm{E}\big[|x|^k + (\mathrm{E}(|x|))^k\big]}{k!} \\
\leq \quad & 1 + \sum_{k=2}^{\infty} \frac{t^k 2^k \mathrm{E}\big[|x|^k\big]}{k!} \\
\leq \quad & 1 + \sum_{k=2}^{\infty} \frac{(2t)^k (2\zeta_0^2)^{k/2} k \Gamma(k/2)}{k!} \\
\leq \quad & 1 + \sum_{k=2}^{\infty} (4e\zeta_0 t)^k \\
= \quad & 1 + 16e^2\zeta_0^2 t^2 \sum_{k=0}^{\infty} (4e\zeta_0 t)^k \\
\leq \quad & 1 + 32e^2\zeta_0^2 t^2 \quad \text{for } |t| \leq \frac{1}{8e\zeta_0} \\
\leq \quad & \exp(32e^2\zeta_0^2 t^2),
\end{aligned}$$

where the second last inequality follows by noting $\Gamma(k/2) \leq (k/2)^{k/2}$ and $k^{1/k} \leq e^{1/e}$ for any $k \geq 2$, and by Stirling's approximation $k! \geq (k/e)^k$. Hence $|x| - \mathrm{E}(|x|) \sim \mathrm{SE}(64e^2\zeta_0^2, 8e\zeta_0)$. Applying the Bernstein's inequality for subexponential random variables, we have

$$P\Big(n^{-1}\Big|\sum_{i=1}^{n}(|x_{ij}| - \mathrm{E}(|x_{ij}|))\Big| > t\Big) \leq 2\exp\Big\{-\frac{n}{2}\min\big(\frac{t^2}{64e^2\zeta_0^2}, \frac{t}{8e\zeta_0}\big)\Big\}, \tag{D.4}$$

$\forall\ t > 0$. Taking $t = 16e\zeta_0\sqrt{\log p/n}$ in (D.4) and noting that we assume $\log p \leq n/4$, by the union bound, we have

$$P\Big(\max_{1\leq j\leq p} n^{-1}\Big|\sum_{i=1}^{n}(|x_{ij}| - \mathrm{E}(|x_{ij}|)\Big| > 16e\zeta_0\sqrt{\log p/n}\Big) \leq 2\exp(-\log p).$$

Hence, with probability at least $1 - 2\exp(-\log p)$,

$$\sum_{i=1}^{n} |x_{ij}| \geq E(|x_{ij}|) - 16e\zeta_0\sqrt{\log p/n}, \text{ for all } j = 1, \ldots, p.$$

As $\min_{1\leq j\leq p} \mathrm{E}(|x_{ij}|) \geq \zeta_* > 0$ and $32e\zeta_0\sqrt{\log p/n} \leq \zeta_*$, the conclusion of the lemma follows. $\square$

Proof of Lemma 1. This result follows from the generalized KKT condition of the convex difference program. We provide a self-contained derivation below.

(i) The result for the global minimum of QR-LASSO follows directly from the definition of subgradient. To see this, let $\widehat{\beta} = \arg\min_\beta\{Q_n(\beta) + \lambda|\beta_-|\}$ denote the global minimum of QR-LASSO. Then

$$\left\{Q_n(\beta) + \lambda|\beta_-|\right\} - \left\{Q_n(\widehat{\beta}) + \lambda|\widehat{\beta}_-|\right\} \geq 0_p^T(\beta - \widehat{\beta}) = 0,$$

where 0_p denotes a p-dimensional zero vector. Hence, by the definition of subgradient, $0_p \in \partial\{Q_n(\widehat{\beta}) + \lambda|\widehat{\beta}_-|\}$.

(ii) Now consider the case of QR-NCP. Let $\widehat{\beta}$ denote a local minimum of $L_n(\beta) = \tilde{L}_n(\beta) - H(\beta)$, where $\tilde{L}_n(\beta) = Q_n(\beta) + \lambda\sum_{j=2}^p|\beta_j|$ and $H(\beta) = \sum_{j=2}^p h_\lambda(\beta_j)$.Then there exists a neighborhood U of $\widehat{\beta}$ such that $L_n(\beta) \geq L_n(\beta)$, for all $\beta \in U$. Hence $\forall\ \beta \in U$

$$\tilde{L}_n(\beta) - \tilde{L}_n(\widehat{\beta}) \geq H(\beta) - H(\widehat{\beta}) \geq (H'(\widehat{\beta}))^T(\beta - \widehat{\beta}), \tag{D.5}$$

where $H'(\widehat{\beta}) = (0, h'_\lambda(\widehat{\beta}_2), \ldots, h'_\lambda(\widehat{\beta}_p))^T$, and the second inequality follows because $H(\beta)$ is convex and differentiable. The convexity of L_n and (D.5) implies that $H'(\widehat{\beta}) \in \partial\tilde{L}_n(\widehat{\beta})$. This finishes the proof. $\square$

Proof of Lemma 2 (QR-LASSO). The proof of the result under hard sparsity was given in Belloni and Chernozhukov (2011). We include it here for completeness and to facilitate the proof for the result under soft sparsity. Consider the event $\Lambda_n = \left\{\lambda \geq 2||\widetilde{S}_n||_\infty\right\}$, where λ

satisfies the conditions of Lemma 2. Lemma 3 ensures that $P(\Lambda_n) \geq 1 - \delta_n - 2\exp(-\log p)$. Recall $\widehat{v} = \widehat{\beta} - \beta^*$. By the definition of $\widehat{\beta}$, $Q_n(\widehat{\beta}) + \lambda||\widehat{\beta}_-||_1 \leq Q_n(\beta^*) + \lambda||\beta^*_-||_1$. This implies

$$Q_n(\widehat{\beta}) - Q_n(\beta^*) \leq \lambda\big(||\beta^*_-||_1 - ||\beta^*_- + \widehat{v}_-||_1\big) \leq \lambda\big(||\widehat{v}_S||_1 - ||\widehat{v}_{S^c}||_1\big). \tag{D.6}$$

On the other hand, the convexity of $Q_n(\cdot)$ guarantees that on Λ_n,

$$Q_n(\widehat{\beta}) - Q_n(\beta^*) \ \geq \ \widetilde{S}_n\widehat{v} \geq -||\widehat{v}||_1||\widetilde{S}_n||_\infty \geq -\frac{\lambda}{2}\big(||\widehat{v}_S||_1 + ||\widehat{v}_{S^c}||_1\big). \tag{D.7}$$

Putting (D.6) and (D.7) together, we have $\widehat{v} \in \Gamma_H$.

Under the soft sparsity assumption, by similar argument as above for (D.6) and (D.7), we obtain that on Λ_n,

$$2\big(||\beta^*_- + \widehat{v}_-||_1 - ||\beta^*_-||_1\big) \leq ||\widehat{v}_{S_a}||_1 + ||\widehat{v}_{S_a^c}||_1, \tag{D.8}$$

where $S_{-a} = \{j : |\beta^*_j| > a, 2 \leq j \leq p\}$ and $S_a = S_{-a} \cup \{1\}$. On the other hand,

$$\begin{aligned} & ||\beta^*_- + \widehat{v}_-||_1 - ||\beta^*_-||_1 \\ \geq \ & \big(||\beta^*_{S_{-a}} + \widehat{v}_{S_a^c}||_1 - ||\beta^*_{S_a^c} + \widehat{v}_{S_{-a}}||_1\big) - \big(||\beta^*_{S_{-a}}||_1 + ||\beta^*_{S_a^c}||_1\big) \\ = \ & ||\widehat{v}_{S_a^c}||_1 - ||\widehat{v}_{S_{-a}}||_1 - 2||\beta^*_{S_a^c}||_1 \\ \geq \ & ||\widehat{v}_{S_a^c}||_1 - ||\widehat{v}_{S_a}||_1 - 2||\beta^*_{S_a^c}||_1. \end{aligned} \tag{D.9}$$

Combining (D.8) with (D.9), we have

$$2\big(||\widehat{v}_{S_a^c}||_1 - ||\widehat{v}_{S_a}||_1 - 2||\beta^*_{S_a^c}||_1\big) \leq ||\widehat{v}_{S_a}||_1 + ||\widehat{v}_{S_a^c}||_1.$$

Hence $\widehat{v} \in \Gamma_W$ under the soft sparsity assumption. $\square$

Proof of Lemma 3 Let $A_{n1} = \{\max_{1\leq j\leq p} \widehat{\sigma}_j^2 \leq m_x\}$, where $\widehat{\sigma}_j^2 = n^{-1}\sum_{i=1}^n x_{ij}^2$. Then

$P(A_{n1}) \geq 1 - \delta_n$, where $\delta_n \to 0$ as $n \to \infty$ by condition(C3). One the event A_{n1}, by the union bound, we have

$$
\begin{aligned}
P\Big(\|\widetilde{S}_n\|_\infty > 2\sqrt{m_x \log p/n}\Big) &= P\Big(\max_{1\leq j\leq p} \big|n^{-1}\sum_{i=1}^n x_{ij}\xi_i\big| > 2\sqrt{m_x \log p/n}\Big) \\
&\leq \sum_{j=1}^p P\Big(\big|n^{-1}\sum_{i=1}^n x_{ij}\xi_i\big| > 2\sqrt{m_x \log p/n}\Big).
\end{aligned}
$$

As $-\tau \leq \xi_i \leq 1-\tau$, ξ is a sub-Gaussian random variable with parameter bounded by one. Hence,

$$
\begin{aligned}
& P\Big(\big|n^{-1}\sum_{i=1}^n x_{ij}\xi_i\big| > 2\sqrt{m_x \log p/n}\Big) \\
\leq\ & 2\mathrm{E}_X\Big\{\exp\big(-\frac{4m_x \log p}{2n^{-1}\sum_{i=1}^n x_{ij}^2}\big)\Big\} \leq 2\exp\big(-2\log p\big).
\end{aligned}
$$

We have

$$
P\Big(\|\widetilde{S}_n\|_\infty > 2\sqrt{m_x \log p/n}\Big) \leq 2\exp\Big(\log p - 2\log p\Big) = 2\exp(-\log p).
$$

This proves the lemma. $\square$

Proof of Corollary 1 (QR-NCP) The geometric structures of the local solutions for non-convex regularized quantile regression are implied by the derivation in the proof of Theorem 3.

For the hard sparsity case, it follows from the first inequality of (C.35) that

$$
\begin{aligned}
\big(a^* - 3\gamma_0/4\big)\|\widehat{v}\|_2^2 &\leq \frac{3}{2}p_\lambda(\widehat{v}_{S_-}) - \frac{1}{2}p_\lambda(\widehat{v}_{S^c}) + \frac{1}{2}p_\lambda(\widehat{v}_1) \\
&\leq \frac{3}{2}p_\lambda(\widehat{v}_S) - \frac{1}{2}p_\lambda(\widehat{v}_{S^c}) \\
&\leq \frac{3}{2}p_\lambda(\widehat{v}_A) - \frac{1}{2}p_\lambda(\widehat{v}_{A^c}),
\end{aligned}
$$

where A is the index set of the largest s elements of $\widehat{v}$ in magnitude. This implies $3p_\lambda(\widehat{v}_A) - p_\lambda(\widehat{v}_{A^c}) \geq 0$. Lemma 5 of Loh and Wainwright (2015) implies that $0 \leq 3p_\lambda(\widehat{v}_A) - p_\lambda(\widehat{v}_{A^c}) \leq$

$\lambda\big(3||\widehat{v}_A||_1 - ||\widehat{v}_{A^c}||_1\big)$ or $||\widehat{v}_{A^c}||_1 \leq 3||\widehat{v}_A||_1$.

For the soft sparsity case, it follows from (C.39) that

$$\frac{1}{2}\lambda||\widehat{v}_{S_a^c}||_1 \;\; \leq \;\; \frac{3}{2}\lambda||\widehat{v}_{S_a}||_1 + \lambda||\beta^*_{S_a^c}||_1,$$

or $||\widehat{v}_{S_a^c}||_1 \leq 3||\widehat{v}_{S_a}||_1 + 2||\beta^*_{S_a^c}||_1$. $\square$

Proof of Corollary 2. The results follow immediately by combining Theorem 4 with the L_1 estimation error bound derived in Theorem 2 and Theorem 3. $\square$

References

Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563.

Arellano, M. and Bonhomme, S. (2017). Quantile selection models with an application to understanding changes in wage inequality. *Econometrica*, 85(1):1–28.

Belloni, A. and Chernozhukov, V. (2011). L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39:82–130.

Belloni, A., Chernozhukov, V., and Kato, K. (2014). Uniform post-selection inference for least absolute deviation regression and other z-estimation problems. *Biometrika*, 102(1):77–94.

Belloni, A., Chernozhukov, V., and Kato, K. (2019). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association*, 114(526):749–758.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bradic, J., Fan, J., and Wang, W. (2011). Penalized composite quasi-likelihood for ultra-high dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):325–349.

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the usa: a quantile regression approach. *Journal of applied econometrics*, 13(1):1–30.

Buchinsky, M. et al. (1994). Changes in the us wage structure 1963-1987: Application of quantile regression. *ECONOMETRICA-EVANSTON ILL-*, 62:405–405.

Bunea, F., Tsybakov, A., Wegkamp, M., et al. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194.

Chamberlain, G. (1994). Quantile regression, censoring, and the structure of wages. In *Advances in econometrics: sixth world congress*, volume 2, pages 171–209.

Chen, X., Li, D., Li, Q., and Li, Z. (2019a). Nonparametric estimation of conditional quantile functions in the presence of irrelevant covariates. *Journal of Econometrics*, 212(2):433–450.

Chen, X., Liu, W., Zhang, Y., et al. (2019b). Quantile regression under memory constraint. *The Annals of Statistics*, 47(6):3244–3273.

Chernozhukov, V. and Fernández-Val, I. (2011). Inference for extremal conditional quantile models, with an application to market and birthweight risks. *The Review of Economic Studies*, 78(2):559–589.

Chernozhukov, V., Fernández-Val, I., Hahn, J., and Newey, W. (2013). Average and quantile effects in nonseparable panel models. *Econometrica*, 81(2):535–580.

Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk overl p-balls forl p-error. *Probability Theory and Related Fields*, 99(2):277–303.

Elsener, A. and van de Geer, S. (2018). Sharp oracle inequalities for stationary points of non-convex penalized m-estimators. *IEEE Transactions on Information Theory*, 65(3):1452–1472.

Fan, J., Fan, Y., and Barut, E. (2014). Adaptive robust variable selection. *Annals of statistics*, 42(1):324.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle property. *Journal of the American Statistical Association*, 96:1348–1360.

Fan, Z. and Lian, H. (2018). Quantile regression for additive coefficient models in high dimensions. *Journal of Multivariate Analysis*, 164:54–64.

Firpo, S., Fortin, N. M., and Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, 77(3):953–973.

Fitzenberger, B., Koenker, R., and Machado, J. A. (2013). *Economic applications of quantile regression*. Springer Science & Business Media.

Galvao, A. F., Lamarche, C., and Lima, L. R. (2013). Estimation of censored quantile regression for panel data with fixed effects. *Journal of the American Statistical Association*, 108(503):1075–1089.

Graham, B. S., Hahn, J., Poirier, A., and Powell, J. L. (2018). A quantile correlated random coefficients panel data model. *Journal of Econometrics*, 206(2):305–335.

Greenshtein, E., Ritov, Y., et al. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Harding, M. and Lamarche, C. (2018). A panel quantile approach to attrition bias in big data: Evidence from a randomized experiment. *Journal of Econometrics*.

Honda, T., Ing, C.-K., Wu, W.-Y., et al. (2019). Adaptively weighted group lasso for semiparametric quantile regression models. *Bernoulli*, 25(4B):3311–3338.

Horowitz, J. L. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 100(472):1238–1249.

Horowitz, J. L. and Spokoiny, V. G. (2002). An adaptive, rate-optimal test of linearity for median regression models. *Journal of the American Statistical Association*, 97(459):822–835.

Kai, B., Li, R., and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of Statistics*, 39:305–332.

Kato, K. (2011). Group lasso for high dimensional sparse quantile regression models. *arXiv preprint arXiv:1103.1458*.

Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics*, 9:155–176.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Koenker, R., Chernozhukov, V., He, X., and Peng, L., editors (2017). *Handbook of Quantile Regression*. Chapman and Hall/CRC.

Koenker, R. and Xiao, Z. (2006). Quantile autoregression. *Journal of the American Statistical Association*, 101(475):980–990.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Lee, E. R., Noh, H., and Park, B. U. (2014). Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229.

Lee, S., Liao, Y., Seo, M. H., and Shin, Y. (2018). Oracle estimation of a change point in high dimensional quantile regression. *Journal of the American Statistical Association*, 43:1184–1194.

Li, Y. J. and Zhu, J. (2008). L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17:163 – 185.

Linton, O. B. and Whang, Y.-J. (2004). A quantilogram approach to evaluating directional predictability.

Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616.

Lv, S., Lin, H., Lian, H., Huang, J., et al. (2018). Oracle inequalities for sparse additive quantile regression in reproducing kernel hilbert space. *The Annals of Statistics*, 46(2):781–813.

Mei, S., Bai, Y., Montanari, A., et al. (2018). The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Nolan, J. (2003). *Stable distributions: models for heavy-tailed data*. Birkhauser New York.

Park, S., He, X., and Zhou, S. (2017). Dantzig-type penalization for multiple quantile regression with high dimensional covariates. *Statistica Sinica*, pages 1619–1638.

Raskutti, G., Wainwright, M. J., and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE transactions on information theory*, 57(10):6976–6994.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, 75(372):828–838.

Sherwood, B., Wang, L., et al. (2016). Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics*, 44(1):288–317.

Shows, J. H., Lu, W., and Zhang, H. H. (2010). Sparse estimation and inference for censored median regression. *Journal of Statistical Planning and Inference*, 140:1903–1917.

Su, L. and Hoshino, T. (2016). Sieve instrumental variable quantile regression estimation of functional coefficient models. *Journal of Econometrics*, 191(1):231–254.

Tang, Y., Song, X., Wang, H. J., and Zhu, Z. (2013). Variable selection in high-dimensional quantile varying coefficient models. *Journal of Multivariate Analysis*, 122:115–132.

Tao, P. D. and An, L. (1997). Convex analysis approach to dc programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288.

van de Geer, S. A. (2000). *Empirical Processes in M-estimation*. Cambridge university press.

van de Geer, S. A. (2016). *Estimation and Testing under Sparsity*. Springer.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media.

Wagener, J., Volgushev, S., and Dette, H. (2012). The quantile process under random censoring. *Mathematical Methods of Statistics*, 21:127–141.

Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25:347–355.

Wang, H., Zhou, J., and Li, Y. (2013a). Variable selection for censored quantile regression. *Statistica Sinica*, 23:145–167.

Wang, L. (2013). The l1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151.

Wang, L., Kim, Y., and Li, R. (2013b). Calibrating non-convex penalized regression in ultra-high dimension. *Annals of statistics*, 41(5):2505.

Wang, L., Wu, Y., and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222.

Wu, Y. C. and Liu, Y. F. (2009). Variable selection in quantile regression. *Statistica Sinica*, 19:801–817.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942.

Zhao, T., Kolar, M., and Liu, H. (2014). A general framework for robust testing and confidence regions in high-dimensional quantile regression. *arXiv preprint arXiv:1412.8724*.

Zheng, Q., Peng, L., and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *Annals of statistics*, 43(5):2225.

Zhong, W., Zhu, L., Li, R., and Cui, H. (2016). Regularized quantile regression and robust feature screening for single index models. *Statistica Sinica*, 26(1):69.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, 36:1108–1126.